1. Sampling and Data
    1. Sampling and Data: Introduction
    2. Definitions of Statistics, Probability, and Key Terms
    3. Data, Sampling, and Variation
    4. Frequency and Frequency Tables
    5. Experimental Design and Ethics
2. Descriptive Statistics
    1. Descriptive Statistics: Introduction
    2. Stem-and-Leaf Graphs (Stemplots), Line Graphs, and Bar Graphs
    3. Histograms, Frequency Polygons, and Time Series Graphs
    4. Measures of the Location of the Data
    5. Box Plots
    6. Measures of the Center of the Data
    7. Skewness and the Mean, Median, and Mode
    8. Measuring the Spread of the Data
3. Probability Topics
    1. Probability Topics: Introduction
    2. Terminology
    3. Independent and Mutually Exclusive Events
    4. Two Basic Rules of Probability
    5. Contingency Tables
    6. Tree and Venn Diagrams (Optional)
4. Discrete Random Variables
    1. Discrete Random Variables: Introduction
    2. Probability Distribution Function (PDF) for a Discrete Random Variable
    3. Mean or Expected Value and Standard Deviation
    4. The Binomial Distribution
    5. The Poisson Distribution
5. Continuous Random Variables

    1. Continuous Random Variables: Introduction
    2. Continuous Probability Functions
    3. The Uniform Distribution
    4. The Exponential Distribution (Optional)
6. Normal Distributions
    1. The Normal Distribution: Introduction
    2. The Standard Normal Distribution
    3. Using the Normal Distribution
7. The Central Limit Theorem
    1. The Central Limit Theorem: Introduction
    2. The Central Limit Theorem for Sample Means (Averages)
    3. Using the Central Limit Theorem
8. Confidence Intervals
    1. Confidence Intervals: Introduction
    2. Estimating a Single Population Mean using the Normal Distribution
    3. Estimating a Single Population Mean using the Student t-Distribution
    4. Estimating a Population Proportion
9. Hypothesis Testing with One Sample
    1. Hypothesis Testing With One Sample: Introduction
    2. The Null and Alternative Hypotheses
    3. Outcomes and the Type I and Type II Errors
    4. Distribution Needed for Hypothesis Testing
    5. Rare Events, the Sample, Decision, and Conclusion
    6. Additional Information and Full Hypothesis Test Examples
10. Hypothesis Testing with Two Samples
    1. Hypothesis Testing With Two Samples: Introduction
    2. Two Population Means with Unknown Standard Deviations
    3. Comparing Two Independent Population Proportions

    4. Matched or Paired Samples
11. The Chi-Square Distribution
    1. The Chi Square Distribution: Introduction
    2. Facts About the Chi-Square Distribution
    3. Goodness-of-Fit Test
    4. Test of Independence
12. Linear Regression and Correlation
    1. Linear Regression and Correlation: Introduction
    2. Linear Equations
    3. Scatterplots
    4. The Regression Equation
    5. Prediction
    6. Testing the Significance of the Correlation Coefficient (Optional)
13. Solutions Sheets
14. Tables
15. Data Sets
16. Notes for the TI-83, 83+, 84, and 84+ Calculators

Sampling and Data: Introduction

class="introduction"

Introduction

We encounter statistics in our daily lives more often than we probably realize and from many different sources, like the news. (credit: David Sim)

Note:

Chapter Objectives

By the end of this chapter, the student should be able to:

- Recognize and differentiate between key terms.
- Apply various types of sampling methods to data collection.
- Create and interpret frequency tables.

You are probably asking yourself the question, "When and where will I use statistics?" If you read any newspaper, watch television, or use the Internet, you will see statistical information. There are statistics about crime, sports, education, politics, and real estate. Typically, when you read a newspaper article or watch a television news program, you are given sample information. With this information, you may make a decision about the correctness of a statement, claim, or "fact." Statistical methods can help you make the "best educated guess."

Since you will undoubtedly be given statistical information at some point in your life, you need to know some techniques for analyzing the information thoughtfully. Think about buying a house or managing a budget. Think about your chosen profession. The fields of economics, business, psychology, education, biology, law, computer science, police science, and early childhood development require at least one course in statistics.

Included in this chapter are the basic ideas and words of probability and statistics. You will soon understand that statistics and probability work together. You will also learn how data are gathered and what "good" data can be distinguished from "bad."

Definitions of Statistics, Probability, and Key Terms

The science of **statistics** deals with the collection, analysis, interpretation, and presentation of **data**. We see and use data in our everyday lives.

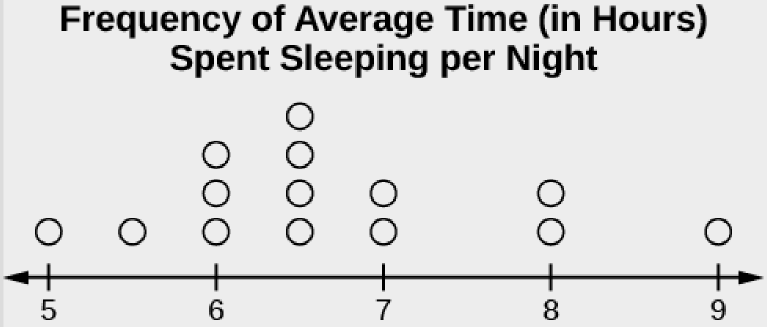
Note:

Collaborative Exercise

In your classroom, try this exercise. Have class members write down the average time (in hours, to the nearest half-hour) they sleep per night. Your instructor will record the data. Then create a simple graph (called a **dot plot**) of the data. A dot plot consists of a number line and dots (or points) positioned above the number line. For example, consider the following data:

5 5.5 6 6 6 6.5 6.5 6.5 6.5 7 7 8 8 9

The dot plot for this data would be as follows:



Does your dot plot look the same as or different from the example? Why? If you did the same example in an English class with the same number of students, do you think the results would be the same? Why or why not?

Where do your data appear to cluster? How might you interpret the clustering?

The questions above ask you to analyze and interpret your data. With this example, you have begun your study of statistics.

In this course, you will learn how to organize and summarize data. Organizing and summarizing data is called **descriptive statistics**. Two ways to summarize data are by graphing and by using numbers (for example, finding an average). After you have studied probability and probability distributions, you will use formal methods for drawing conclusions from "good" data. The formal methods are called **inferential statistics**. Statistical inference uses probability to determine how confident we can be that our conclusions are correct.

Effective interpretation of data (inference) is based on good procedures for producing data and thoughtful examination of the data. You will encounter what will seem to be too many mathematical formulas for interpreting data. The goal of statistics is not to perform numerous calculations using the formulas, but to gain an understanding of your data. The calculations can be done using a calculator or a computer. The understanding must come from you. If you can thoroughly grasp the basics of statistics, you can be more confident in the decisions you make in life.

Probability

Probability is a mathematical tool used to study randomness. It deals with the chance (the likelihood) of an event occurring. For example, if you toss a **fair** coin four times, the outcomes may not be two heads and two tails. However, if you toss the same coin 4,000 times, the outcomes will be close to half heads and half tails. The expected theoretical probability of heads in any one toss is $\frac{1}{2}$ or 0.5. Even though the outcomes of a few repetitions are uncertain, there is a regular pattern of outcomes when there are many repetitions. After reading about the English statistician Karl **Pearson** who tossed a coin 24,000 times with a result of 12,012 heads, one of the authors tossed a coin 2,000 times. The results were 996 heads. The fraction $\frac{996}{2000}$ is equal to 0.498 which is very close to 0.5, the expected probability.

The theory of probability began with the study of games of chance such as poker. Predictions take the form of probabilities. To predict the likelihood of an earthquake, of rain, or whether you will get an A in this course, we use probabilities. Doctors use probability to determine the chance of a vaccination causing the disease the vaccination is supposed to prevent. A stockbroker uses probability to determine the rate of return on a client's investments. You might use probability to decide to buy a lottery ticket or not. In your study of statistics, you will use the power of mathematics through probability calculations to analyze and interpret your data.

Key Terms

In statistics, we generally want to study a **population**. You can think of a population as a collection of persons, things, or objects under study. To study the population, we select a **sample**. The idea of **sampling** is to select a portion (or subset) of the larger population and study that portion (the sample) to gain information about the population. Data are the result of sampling from a population.

Because it takes a lot of time and money to examine an entire population, sampling is a very practical technique. If you wished to compute the overall grade point average at your school, it would make sense to select a sample of students who attend the school. The data collected from the sample would be the students' grade point averages. In presidential elections, opinion poll samples of 1,000–2,000 people are taken. The opinion poll is supposed to represent the views of the people in the entire country. Manufacturers of canned carbonated drinks take samples to determine if a 16 ounce can contains 16 ounces of carbonated drink.

From the sample data, we can calculate a statistic. A **statistic** is a number that represents a property of the sample. For example, if we consider one math class to be a sample of the population of all math classes, then the average number of points earned by students in that one math class at the end of the term is an example of a statistic. The statistic is an estimate of a population parameter. A **parameter** is a number that is a property of the population. Since we considered all math classes to be the population, then the average number of points earned per student over all the math classes is an example of a parameter.

One of the main concerns in the field of statistics is how accurately a statistic estimates a parameter. The accuracy really depends on how well the sample represents the population. The sample must contain the characteristics of the population in order to be a **representative sample**. We are interested in both the sample statistic and the population parameter in inferential statistics. In a later chapter, we will use the sample statistic to test the validity of the established population parameter.

A **variable**, notated by capital letters such as X and Y, is a characteristic of interest for each person or thing in a population. Variables may be **numerical** or **categorical**. **Numerical variables** take on values with equal units such as weight in pounds and time in hours. **Categorical variables** place the person or thing into a category. If we let X equal the number of points earned by one math student at the end of a term, then X is a numerical variable. If we let Y be a person's party affiliation, then some examples of Y include Republican, Democrat, and Independent. Y is a categorical variable. We could do some math with values of X (calculate the average number of points earned, for example), but it makes no sense to do math with values of Y (calculating an average party affiliation makes no sense).

Data are the actual values of the variable. They may be numbers or they may be words. **Datum** is a single value.

Two words that come up often in statistics are **mean** and **proportion**. If you were to take three exams in your math classes and obtain scores of 86, 75, and 92, you would calculate your mean score by adding the three exam scores and dividing by three (your mean score would be 84.3 to one decimal place). If, in your math class, there are 40 students and 22 are men and 18 are women, then the proportion of men students is $\frac{22}{40}$ and the proportion of women students is $\frac{18}{40}$. Mean and proportion are discussed in more detail in later chapters.

Note:
NOTE
The words "**mean**" and "**average**" are often used interchangeably. The substitution of one word for the other is common practice. The technical term is "arithmetic mean," and "average" is technically a center location. However, in practice among non-statisticians, "average" is commonly accepted for "arithmetic mean."

Example:
Exercise:

Problem:

Determine what the key terms refer to in the following study. We want to know the average (mean) amount of money first year college students spend at ABC College on school supplies that do not include books. We randomly survey 100 first year students at the college. Three of those students spent $150, $200, and $225, respectively.

Solution:

The **population** is all first year students attending ABC College this term.

The **sample** could be all students enrolled in one section of a beginning statistics course at ABC College (although this sample may not represent the entire population).

The **parameter** is the average (mean) amount of money spent (excluding books) by first year college students at ABC College this term.

The **statistic** is the average (mean) amount of money spent (excluding books) by first year college students in the sample.

The **variable** could be the amount of money spent (excluding books) by one first year student. Let X = the amount of money spent (excluding books) by one first year student attending ABC College.

The **data** are the dollar amounts spent by the first year students. Examples of the data are \$150, \$200, and \$225.

Note:
Try It
Exercise:

Problem:

Determine what the key terms refer to in the following study. We want to know the average (mean) amount of money spent on school uniforms each year by families with children at Knoll Academy. We randomly survey 100 families with children in the school. Three of the families spent \$65, \$75, and \$95, respectively.

Solution:

The **population** is all families with children attending Knoll Academy.

The **sample** is a random selection of 100 families with children attending Knoll Academy.

The **parameter** is the average (mean) amount of money spent on school uniforms by families with children at Knoll Academy.

The **statistic** is the average (mean) amount of money spent on school uniforms by families in the sample.

The **variable** is the amount of money spent by one family. Let X = the amount of money spent on school uniforms by one family with children attending Knoll Academy.

The **data** are the dollar amounts spent by the families. Examples of the data are \$65, \$75, and \$95.

Example:
Exercise:

Problem:

Determine what the key terms refer to in the following study.

A study was conducted at a local college to analyze the average cumulative GPA’s of students who graduated last year. Fill in the letter of the phrase that best describes each of the items below.

1._____ Population 2._____ Statistic 3._____ Parameter 4._____ Sample 5._____ Variable 6._____ Data

- a) all students who attended the college last year
- b) the cumulative GPA of one student who graduated from the college last year
- c) 3.65, 2.80, 1.50, 3.90
- d) a group of students who graduated from the college last year, randomly selected
- e) the average cumulative GPA of students who graduated from the college last year
- f) all students who graduated from the college last year
- g) the average cumulative GPA of students in the study who graduated from the college last year

Solution:

1. f 2. g 3. e 4. d 5. b 6. c

Example:
Exercise:

Problem:

Determine what the key terms refer to in the following study.

As part of a study designed to test the safety of automobiles, the National Transportation Safety Board collected and reviewed data about the effects of an automobile crash on test dummies. Here is the criterion they used:

Speed at which Cars Crashed	Location of “drive” (i.e. dummies)
35 miles/hour	Front Seat

Cars with dummies in the front seats were crashed into a wall at a speed of 35 miles per hour. We want to know the proportion of dummies in the driver’s seat that would have had head injuries, if they had been actual drivers. We start with a simple random sample of 75 cars.

Solution:

The **population** is all cars containing dummies in the front seat.

The **sample** is the 75 cars, selected by a simple random sample.

The **parameter** is the proportion of driver dummies (if they had been real people) who would have suffered head injuries in the population.

The **statistic** is proportion of driver dummies (if they had been real people) who would have suffered head injuries in the sample.

The **variable** X = the number of driver dummies (if they had been real people) who would have suffered head injuries.

The **data** are either: yes, had head injury, or no, did not.

Example:
Exercise:

Problem:

Determine what the key terms refer to in the following study.

An insurance company would like to determine the proportion of all medical doctors who have been involved in one or more malpractice lawsuits. The company selects 500 doctors at random from a professional directory and determines the number in the sample who have been involved in a malpractice lawsuit.

Solution:

The **population** is all medical doctors listed in the professional directory.

The **parameter** is the proportion of medical doctors who have been involved in one or more malpractice suits in the population.

The **sample** is the 500 doctors selected at random from the professional directory.

The **statistic** is the proportion of medical doctors who have been involved in one or more malpractice suits in the sample.

The **variable** X = the number of medical doctors who have been involved in one or more malpractice suits.

The **data** are either: yes, was involved in one or more malpractice lawsuits, or no, was not.

Note:
Collaborative Exercise
Do the following exercise collaboratively with up to four people per group. Find a population, a sample, the parameter, the statistic, a variable, and data for the following study: You want to determine the average (mean) number of glasses of milk college students drink per day. Suppose yesterday, in your English class, you asked five students how many glasses of milk they drank the day before. The answers were 1, 0, 1, 3, and 4 glasses of milk.

References

The Data and Story Library, http://lib.stat.cmu.edu/DASL/Stories/CrashTestDummies.html (accessed May 1, 2013).

Section Review

The mathematical theory of statistics is easier to learn when you know the language. This module presents important terms that will be used throughout the text.

Practice

Use the following information to answer the next five exercises. Studies are often done by pharmaceutical companies to determine the effectiveness of a treatment program. Suppose that a new AIDS antibody drug is currently under study. It is given to patients once the AIDS symptoms have revealed themselves. Of interest is the average (mean) length of time in months patients live once they start the treatment. Two researchers each follow a different set of 40 patients with AIDS from the start of treatment until their deaths. The following data (in months) are collected.

Researcher A:
3 4 11 15 16 17 22 44 37 16 14 24 25 15 26 27 33 29 35 44 13 21 22 10 12 8 40 32 26 27 31 34 29 17 8 24 18 47 33 34

Researcher B:
3 14 11 5 16 17 28 41 31 18 14 14 26 25 21 22 31 2 35 44 23 21 21 16 12 18 41 22 16 25 33 34 29 13 18 24 23 42 33 29

Determine what the key terms refer to in the example for Researcher A.

Exercise:

Problem: population

Solution:

AIDS patients.

Exercise:

Problem: sample

Exercise:

Problem: parameter

Solution:

The average length of time (in months) AIDS patients live after treatment.

Exercise:

Problem: statistic

Exercise:

Problem: variable

Solution:

X = the length of time (in months) AIDS patients live after treatment

HOMEWORK

For each of the following eight exercises, identify: a. the population, b. the sample, c. the parameter, d. the statistic, e. the variable, and f. the data. Give examples where appropriate.

Exercise:

Problem:

A fitness center is interested in the mean amount of time a client exercises in the center each week.

Exercise:

Problem:

Ski resorts are interested in the mean age that children take their first ski and snowboard lessons. They need this information to plan their ski classes optimally.

Solution:

a. all children who take ski or snowboard lessons
b. a group of these children
c. the population mean age of children who take their first snowboard lesson
d. the sample mean age of children who take their first snowboard lesson
e. X = the age of one child who takes his or her first ski or snowboard lesson
f. values for X, such as 3, 7, and so on

Exercise:

Problem:

A cardiologist is interested in the mean recovery period of her patients who have had heart attacks.

Exercise:

Problem:

Insurance companies are interested in the mean health costs each year of their clients, so that they can determine the costs of health insurance.

Solution:

a. the clients of the insurance companies
b. a group of the clients
c. the mean health costs of the clients
d. the mean health costs of the sample
e. X = the health costs of one client
f. values for X, such as 34, 9, 82, and so on

Exercise:

Problem:

A politician is interested in the proportion of voters in his district who think he is doing a good job.

Exercise:

Problem:

A marriage counselor is interested in the proportion of clients she counsels who stay married.

Solution:

a. all the clients of this counselor
b. a group of clients of this marriage counselor
c. the proportion of all her clients who stay married
d. the proportion of the sample of the counselor's clients who stay married
e. X = the number of couples who stay married
f. yes, no

Exercise:

Problem:

Political pollsters may be interested in the proportion of people who will vote for a particular cause.

Exercise:

Problem:

A marketing company is interested in the proportion of people who will buy a particular product.

Solution:

a. all people (maybe in a certain geographic area, such as the United States)
b. a group of the people
c. the proportion of all people who will buy the product
d. the proportion of the sample who will buy the product
e. X = the number of people who will buy it
f. buy, not buy

Use the following information to answer the next three exercises: A Lake Tahoe Community College instructor is interested in the mean number of days Lake Tahoe Community College math students are absent from class during a quarter.

Exercise:

Problem: What is the population she is interested in?

a. all Lake Tahoe Community College students
b. all Lake Tahoe Community College English students
c. all Lake Tahoe Community College students in her classes
d. all Lake Tahoe Community College math students

Exercise:

Problem: Consider the following:

X = number of days a Lake Tahoe Community College math student is absent

In this case, X is an example of a:

a. variable.
b. population.
c. statistic.
d. data.

Solution:

a

Exercise:

Problem:

The instructor’s sample produces a mean number of days absent of 3.5 days. This value is an example of a:

a. parameter.
b. data.
c. statistic.
d. variable.

Glossary

Average

also called mean; a number that describes the central tendency of the data

Categorical Variable

variables that take on values that are names or labels

Data

a set of observations (a set of possible outcomes); most data can be put into two groups: **qualitative** (an attribute whose value is indicated by a label) or **quantitative** (an attribute whose value is indicated by a number). Quantitative data can be separated into two subgroups: **discrete** and **continuous**. Data is discrete if it is the result of counting (such as the number of students of a given ethnic group in a class or the number of books on a shelf). Data is continuous if it is the result of measuring (such as distance traveled or weight of luggage)

Numerical Variable

variables that take on values that are indicated by numbers

Parameter

a number that is used to represent a population characteristic and that generally cannot be determined easily

Population

all individuals, objects, or measurements whose properties are being studied

Probability

a number between zero and one, inclusive, that gives the likelihood that a specific event will occur

Proportion

the number of successes divided by the total number in the sample

Representative Sample
a subset of the population that has the same characteristics as the population

Sample
a subset of the population studied

Statistic
a numerical characteristic of the sample; a statistic estimates the corresponding population parameter.

Variable
a characteristic of interest for each person or object in a population

Data, Sampling, and Variation

Data may come from a population or from a sample. Small letters like x or y generally are used to represent data values. Most data can be put into the following categories:

- Qualitative
- Quantitative

Qualitative data are the result of categorizing or describing attributes of a population. Hair color, blood type, ethnic group, the car a person drives, and the street a person lives on are examples of qualitative data. Qualitative data are generally described by words or letters. For instance, hair color might be black, dark brown, light brown, blonde, gray, or red. Blood type might be AB+, O-, or B+. Researchers often prefer to use quantitative data over qualitative data because it lends itself more easily to mathematical analysis. For example, it does not make sense to find an average hair color or blood type.

Quantitative data are always numbers. Quantitative data are the result of **counting** or **measuring** attributes of a population. Amount of money, pulse rate, weight, number of people living in your town, and number of students who take statistics are examples of quantitative data. Quantitative data may be either **discrete** or **continuous**.

All data that are the result of counting are called **quantitative discrete data**. These data take on only certain numerical values. If you count the number of phone calls you receive for each day of the week, you might get values such as zero, one, two, or three.

All data that are the result of measuring are **quantitative continuous data** assuming that we can measure accurately. Measuring angles in radians might result in such numbers as $\frac{\pi}{6}$, $\frac{\pi}{3}$, $\frac{\pi}{2}$, π, $\frac{3\pi}{4}$, and so on. If you and your friends carry backpacks with books in them to school, the numbers of books in the backpacks are discrete data and the weights of the backpacks are continuous data.

Example:
Data Sample of Quantitative Discrete Data
The data are the number of books students carry in their backpacks. You sample five students. Two students carry three books, one student carries four books, one student carries two books, and one student carries one book. The numbers of books (three, four, two, and one) are the quantitative discrete data.

Note:
Try It
Exercise:

Problem:

The data are the number of machines in a gym. You sample five gyms. One gym has 12 machines, one gym has 15 machines, one gym has ten machines, one gym has 22 machines, and the other gym has 20 machines. What type of data is this?

Solution:

quantitative discrete data

Example:
Data Sample of Quantitative Continuous Data
The data are the weights of backpacks with books in them. You sample the same five students. The weights (in pounds) of their backpacks are 6.2, 7, 6.8, 9.1, 4.3. Notice that backpacks carrying three books can have different weights. Weights are quantitative continuous data because weights are measured.

Note:
Try It
Exercise:

Problem:

The data are the areas of lawns in square feet. You sample five houses. The areas of the lawns are 144 sq. feet, 160 sq. feet, 190 sq. feet, 180 sq. feet, and 210 sq. feet. What type of data is this?

Solution:

quantitative continuous data

Example:

You go to the supermarket and purchase three cans of soup (19 ounces tomato bisque, 14.1 ounces lentil, and 19 ounces Italian wedding), two packages of nuts (walnuts and peanuts), four different kinds of vegetable (broccoli, cauliflower, spinach, and carrots), and two desserts (16 ounces Cherry Garcia ice cream and 32 ounces chocolate chip cookies).

Exercise:

Problem:

Name data sets that are quantitative discrete, quantitative continuous, and qualitative.

Solution:

One Possible Solution:

- The three cans of soup, two packages of nuts, four kinds of vegetables, and two desserts are quantitative discrete data because you count them.
- The weights of the soups (19 ounces, 14.1 ounces, 19 ounces) are quantitative continuous data because you measure weights as precisely as possible.
- Types of soups, nuts, vegetables, and desserts are qualitative data because they are categorical.

Try to identify additional data sets in this example.

Example:
The data are the colors of backpacks. Again, you sample the same five students. One student has a red backpack, two students have black backpacks, one student has a green backpack, and one student has a gray backpack. The colors red, black, black, green, and gray are qualitative data.

Note:
Try It
Exercise:

Problem:

The data are the colors of houses. You sample five houses. The colors of the houses are white, yellow, white, red, and white. What type of data is this?

Solution:

qualitative data

Note:
Note
You may collect data as numbers and report it categorically. For example, the quiz scores for each student are recorded throughout the term. At the end of the term, the quiz scores are reported as A, B, C, D, or F.

Example:
Exercise:

Problem:

Work collaboratively to determine the correct data type (quantitative or qualitative). Indicate whether quantitative data are continuous or discrete. Hint: Data that are discrete often start with the words "the number of."

a. the number of pairs of shoes you own
b. the type of car you drive
c. where you go on vacation
d. the distance it is from your home to the nearest grocery store
e. the number of classes you take per school year.
f. the tuition for your classes
g. the type of calculator you use
h. movie ratings
i. political party preferences
j. weights of sumo wrestlers
k. amount of money (in dollars) won playing poker
l. number of correct answers on a quiz
m. peoples’ attitudes toward the government
n. IQ scores (This may cause some discussion.)

Solution:

Items a, e, f, k, and l are quantitative discrete; items d, j, and n are quantitative continuous; items b, c, g, h, i, and m are qualitative.

Note:
Try It
Exercise:

Problem:

Determine the correct data type (quantitative or qualitative) for the number of cars in a parking lot. Indicate whether quantitative data are continuous or discrete.

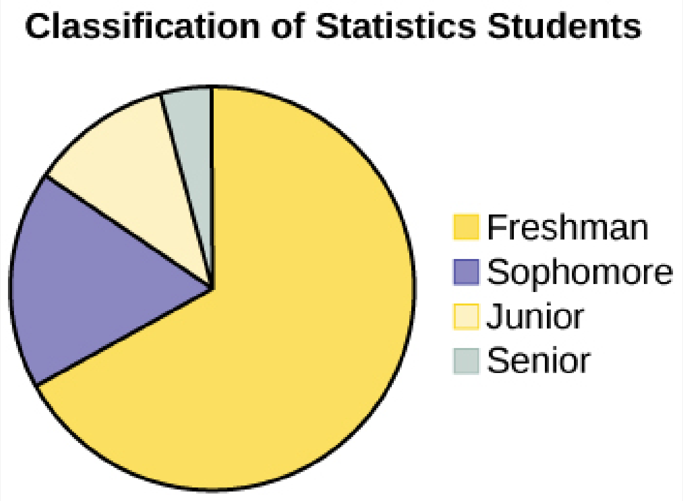
Solution:

quantitative discrete

Example:
Exercise:

Problem:

A statistics professor collects information about the classification of her students as freshmen, sophomores, juniors, or seniors. The data she collects are summarized in the following pie chart. What type of data does this graph show?



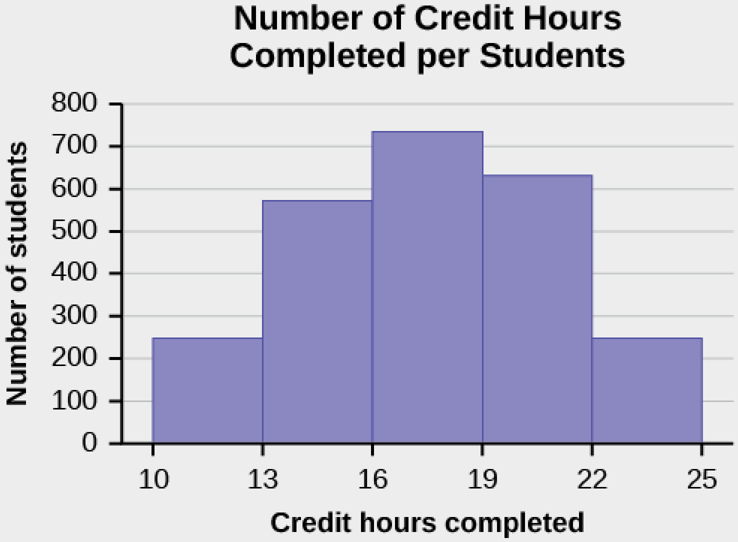
Solution:

This pie chart shows the students in each year, which is **qualitative data**.

Note:
Try It
Exercise:

Problem:

The registrar at State University keeps records of the number of credit hours students complete each semester. The data he collects are summarized in the histogram. The class boundaries are 10 to less than 13, 13 to less than 16, 16 to less than 19, 19 to less than 22, and 22 to less than 25.



What type of data does this graph show?

Solution:

A histogram is used to display quantitative data: the numbers of credit hours completed. Because students can complete only a whole number of hours (no fractions of hours allowed), this data is quantitative discrete.

Qualitative Data Discussion

Below are tables comparing the number of part-time and full-time students at De Anza College and Foothill College enrolled for the spring 2010 quarter. The tables display counts (frequencies) and percentages or proportions (relative frequencies). The percent columns make comparing the same categories in the colleges easier. Displaying percentages along with the numbers is often helpful, but it is particularly important when comparing sets of data that do not have the same totals, such as the total enrollments for both colleges in this example. Notice how much larger the percentage for part-time students at Foothill College is compared to De Anza College.

De Anza College				**Foothill College**		
	Number	Percent			Number	Percent
Full-time	9,200	40.9%		Full-time	4,059	28.6%
Part-time	13,296	59.1%		Part-time	10,124	71.4%
Total	22,496	100%		Total	14,183	100%

Fall Term 2007 (Census day)

Tables are a good way of organizing and displaying data. But graphs can be even more helpful in understanding the data. There are no strict rules concerning which graphs to use. Two graphs that are used to display qualitative data are pie charts and bar graphs.

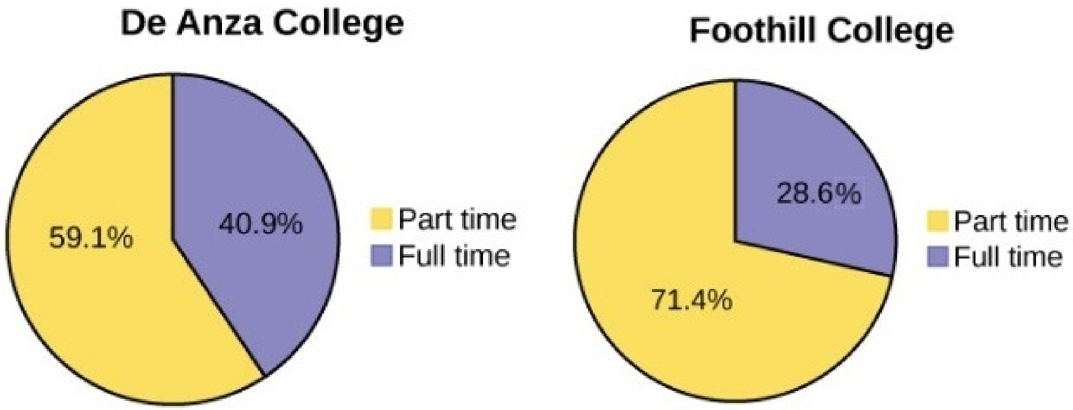
In a **pie chart**, categories of data are represented by wedges in a circle and are proportional in size to the percent of individuals in each category.

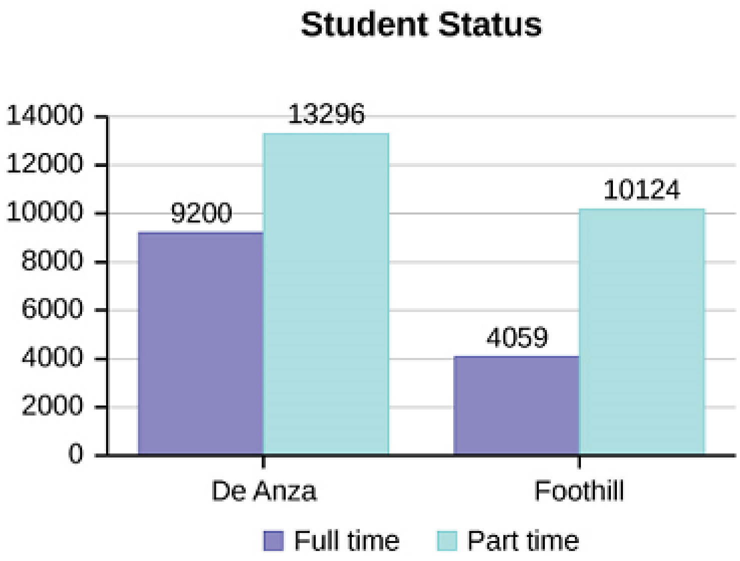
In a **bar graph**, the length of the bar for each category is proportional to the number or percent of individuals in each category. Bars may be vertical or horizontal.

A **Pareto chart** consists of bars that are sorted into order by category size (largest to smallest).

Look at [link] and [link] and determine which graph (pie or bar) you think displays the comparisons better.

It is a good idea to look at a variety of graphs to see which is the most helpful in displaying the data. We might make different choices of what we think is the "best" graph depending on the data and the context. Our choice also depends on what we are using the data for.



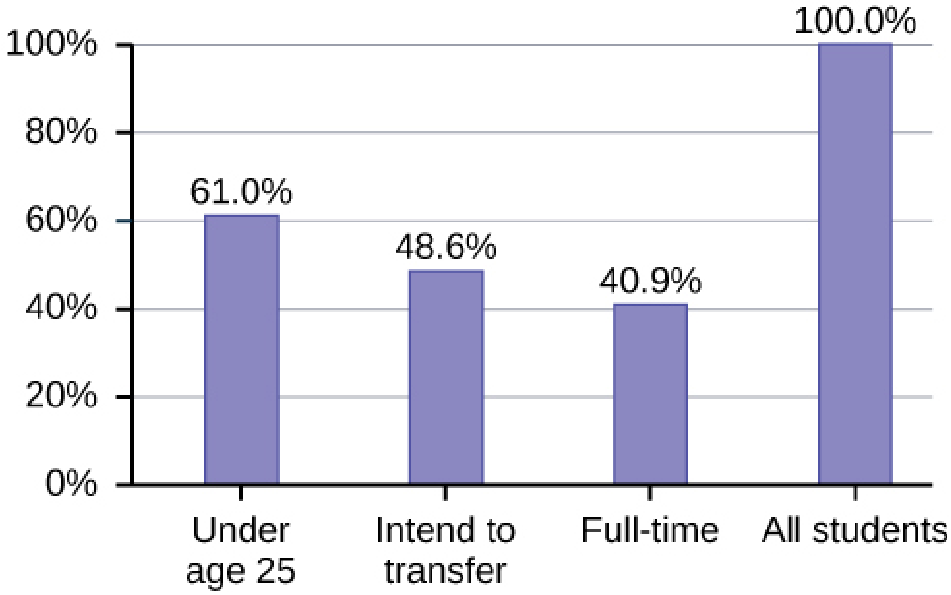


Percentages That Add to More (or Less) Than 100%

Sometimes percentages add up to be more than 100% (or less than 100%). In the graph, the percentages add to more than 100% because students can be in more than one category. A bar graph is appropriate to compare the relative size of the categories. A pie chart cannot be used. It also could not be used if the percentages added to less than 100%.

Characteristic/Category	**Percent**
Full-Time Students	40.9%
Students who intend to transfer to a 4-year educational institution	48.6%
Students under age 25	61.0%
TOTAL	150.5%

De Anza College Spring 2010



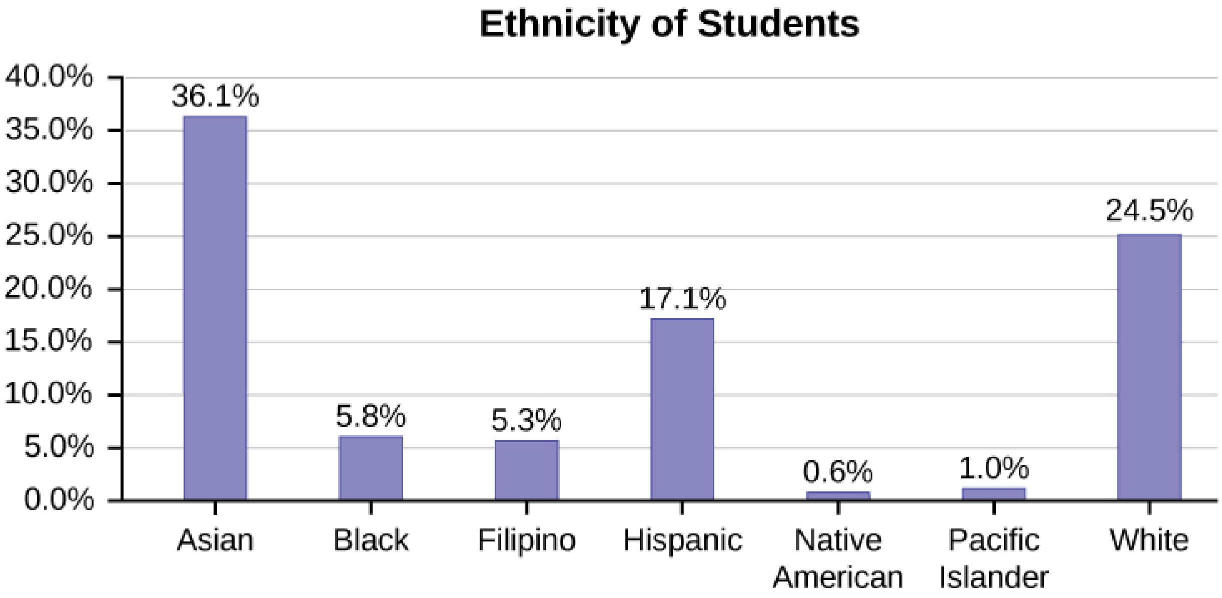
Omitting Categories/Missing Data

The table displays Ethnicity of Students but is missing the "Other/Unknown" category. This category contains people who did not feel they fit into any of the ethnicity categories or declined to respond. Notice that the frequencies do not add up to the total number of students. In this situation, create a bar graph and not a pie chart.

	Frequency	**Percent**
Asian	8,794	36.1%
Black	1,412	5.8%

	Frequency	**Percent**
Filipino	1,298	5.3%
Hispanic	4,180	17.1%
Native American	146	0.6%
Pacific Islander	236	1.0%
White	5,978	24.5%
TOTAL	22,044 out of 24,382	90.4% out of 100%

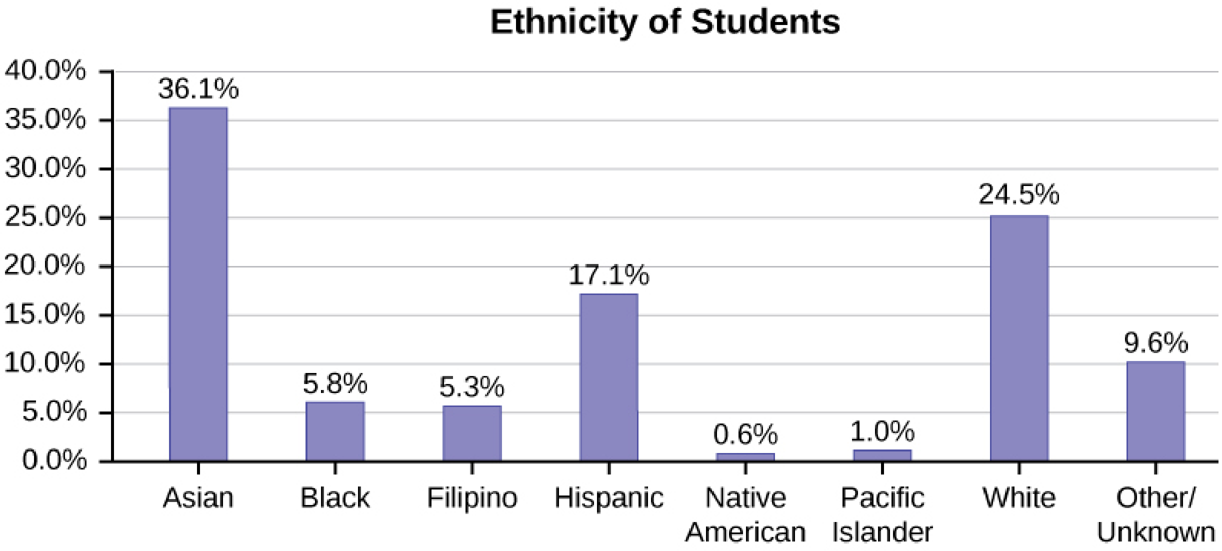
Ethnicity of Students at De Anza College Fall Term 2007 (Census Day)



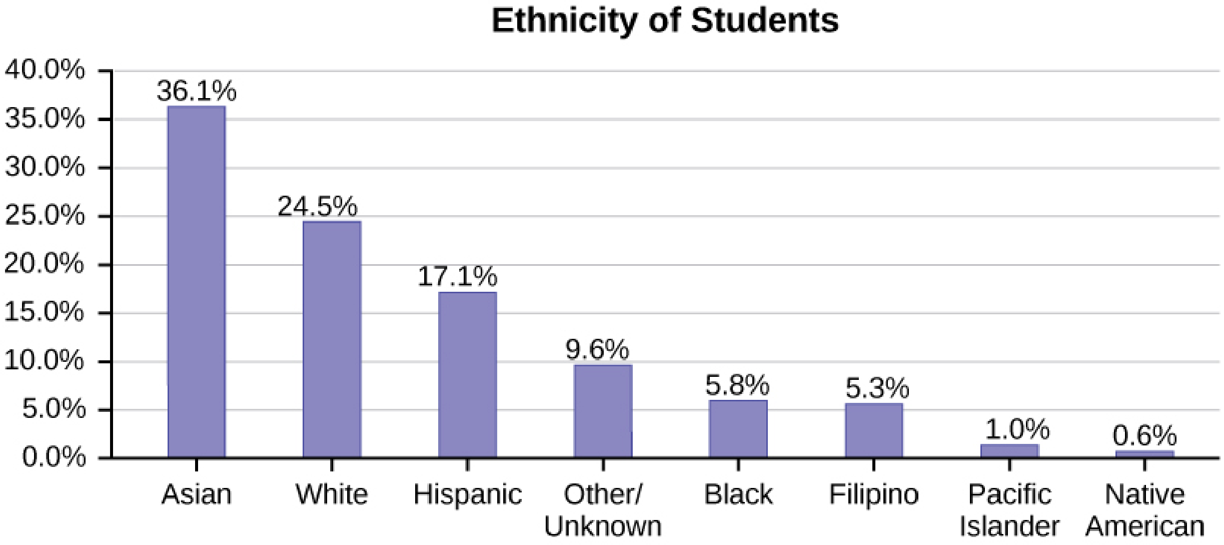
The following graph is the same as the previous graph but the “Other/Unknown” percent (9.6%) has been included. The “Other/Unknown” category is large compared to some of the other categories (Native American, 0.6%, Pacific Islander 1.0%). This is important to know when we think about what the data are telling us.

This particular bar graph in [link] can be difficult to understand visually. The graph in [link] is a Pareto chart. The Pareto chart has the bars sorted from largest to smallest and is easier to read and interpret.

Bar Graph with Other/Unknown Category

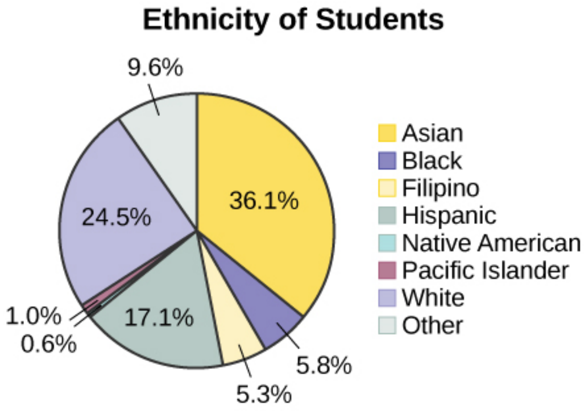


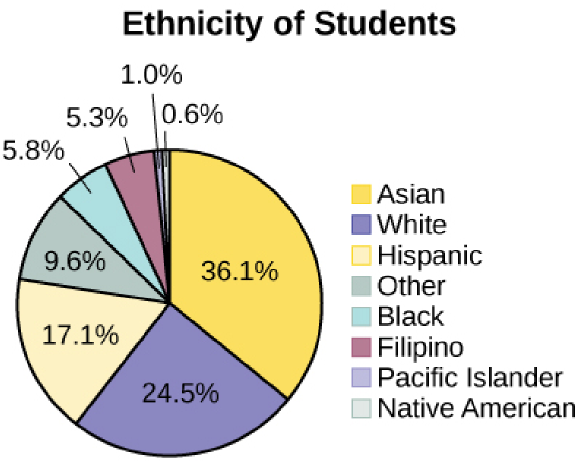
Pareto Chart With Bars Sorted by Size



Pie Charts: No Missing Data

The following pie charts have the “Other/Unknown” category included (since the percentages must add to 100%). The chart in [link] is organized by the size of each wedge, which makes it a more visually informative graph than the unsorted, alphabetical graph in [link].





Sampling

Gathering information about an entire population often costs too much or is virtually impossible. Instead, we use a sample of the population. **A sample should have the same characteristics as the population it is representing.** Most statisticians use various methods of random sampling in an attempt to achieve this goal. This section will describe a few of the most common methods. There are several different methods of **random sampling**. In each form of random sampling, each member of a population initially has an equal chance of being selected for the sample. Each method has pros and cons. The easiest method to describe is called a **simple random sample**. Any group of n individuals is equally likely to be chosen as any other group of n individuals if the simple random sampling technique is used. In other words, each sample of the same size has an equal chance of being selected. For example, suppose Lisa wants to form a four-person study group (herself and three other people) from her pre-calculus class, which has 31 members not including Lisa. To choose a simple random sample of size three from the other members of her class, Lisa could put all 31 names in a hat, shake the hat, close her eyes, and pick out three names. A more technological way is for Lisa to first list the last names of the members of her class together with a two-digit number, as in the table below:

ID	Name	ID	Name	ID	Name
00	Anselmo	11	King	21	Roquero
01	Bautista	12	Legeny	22	Roth
02	Bayani	13	Lundquist	23	Rowell
03	Cheng	14	Macierz	24	Salangsang
04	Cuarismo	15	Motogawa	25	Slade
05	Cuningham	16	Okimoto	26	Stratcher
06	Fontecha	17	Patel	27	Tallai
07	Hong	18	Price	28	Tran
08	Hoobler	19	Quizon	29	Wai
09	Jiao	20	Reyes	30	Wood
10	Khan				

Class Roster

Lisa can use a table of random numbers (found in many statistics books and mathematical handbooks), a calculator, or a computer to generate random numbers. For this example, suppose Lisa chooses to generate random numbers from a calculator. The numbers generated are as follows:

0.94360 0.99832 0.14669 0.51470 0.40581 0.73381 0.04399

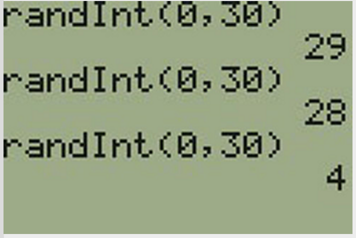
Lisa reads two-digit groups until she has chosen three class members (that is, she reads 0.94360 as the groups 94, 43, 36, 60). Each random number may only contribute one class member. If she needed to, Lisa could have generated more random numbers.

The random numbers 0.94360 and 0.99832 do not contain appropriate two digit numbers. However the third random number, 0.14669, contains 14 (the fourth random number also contains 14), the fifth random number contains 05, and the seventh random number contains 04. The two-digit number 14 corresponds to Macierz, 05 corresponds to Cuningham, and 04 corresponds to Cuarismo. Besides herself, Lisa’s group will consist of Marcierz, Cuningham, and Cuarismo.

Note:
To generate random numbers:

- Press MATH.
- Arrow over to PRB.
- Press 5:randInt(. Enter 0, 30).
- Press ENTER for the first random number.
- Press ENTER two more times for the other 2 random numbers. If there is a repeat press ENTER again.

Note: randInt(0, 30, 3) will generate 3 random numbers.



```
randInt(0,30)
                29
randInt(0,30)
                28
randInt(0,30)
                 4
```

Besides simple random sampling, there are other forms of sampling that involve a chance process for getting the sample. **Other well-known random sampling methods are the stratified sample, the cluster sample, and the systematic sample.**

To choose a **stratified sample**, divide the population into groups called strata and then take a **proportionate** number from each stratum. For example, you could stratify (group) your college population by department and then choose a proportionate simple random sample from each stratum (each department) to get a stratified random sample. To choose a simple random sample from each department, number each member of the first department, number each member of the second department, and do the same for the remaining departments. Then use simple random sampling to choose proportionate numbers from the first department and do the same for each of the remaining departments. Those numbers picked from the first department, picked from the second department, and so on represent the members who make up the stratified sample.

To choose a **cluster sample**, divide the population into clusters (groups) and then randomly select some of the clusters. All the members from these clusters are in the cluster sample. For example, if you randomly sample four departments from your college population, the four departments make up the cluster sample. Divide your college faculty by department. The departments are the clusters. Number each department, and then choose four different numbers using simple random sampling. All members of the four departments with those numbers are the cluster sample.

To choose a **systematic sample**, randomly select a starting point and take every n^{th} piece of data from a listing of the population. For example, suppose you have to do a phone survey. Your phone book contains 20,000 residence listings. You must choose 400 names for the sample. Number the population 1–20,000 and then use a simple random sample to pick a number that represents the first name in the sample. Then choose every fiftieth name thereafter until you have a total of 400 names (you might have to go back to the beginning of your phone list). Systematic sampling is frequently chosen because it is a simple method.

A type of sampling that is non-random is convenience sampling. **Convenience sampling** involves using results that are readily available. For example, a computer software store conducts a marketing study by interviewing potential customers who happen to be in the store browsing through the available software. The results of convenience sampling may be very good in some cases and highly biased (favor certain outcomes) in others.

Sampling data should be done very carefully. Collecting data carelessly can have devastating results. Surveys mailed to households and then returned may be very biased (they may favor a certain group). It is better for the person conducting the survey to select the sample respondents.

True random sampling is done **with replacement**. That is, once a member is picked, that member goes back into the population and thus may be chosen more than once. However for practical reasons, in most populations, simple random sampling is done **without replacement**. Surveys are typically done without replacement. That is, a member of the population may be chosen only once. Most samples are taken from large populations and the sample tends to be small in comparison to the population. Since this is the case, sampling without replacement is approximately the same as sampling with replacement because the chance of picking the same individual more than once with replacement is very low.

In a college population of 10,000 people, suppose you want to pick a sample of 1,000 randomly for a survey. **For any particular sample of 1,000**, if you are sampling **with replacement**,

- the chance of picking the first person is 1,000 out of 10,000 (0.1000);
- the chance of picking a different second person for this sample is 999 out of 10,000 (0.0999);
- the chance of picking the same person again is 1 out of 10,000 (very low).

If you are sampling **without replacement**,

- the chance of picking the first person for any particular sample is 1000 out of 10,000 (0.1000);
- the chance of picking a different second person is 999 out of 9,999 (0.0999);
- you do not replace the first person before picking the next person.

Compare the fractions 999/10,000 and 999/9,999. For accuracy, carry the decimal answers to four decimal places. To four decimal places, these numbers are equivalent (0.0999).

Sampling without replacement instead of sampling with replacement becomes a mathematical issue only when the population is small. For example, if the population is 25 people, the sample is ten, and you are sampling **with replacement for any particular sample**, then the chance of picking the first person is ten out of 25, and the chance of picking a different second person is nine out of 25 (you replace the first person).

If you sample **without replacement**, then the chance of picking the first person is ten out of 25, and then the chance of picking the second person (who is different) is nine out of 24 (you do not replace the first person).

Compare the fractions 9/25 and 9/24. To four decimal places, 9/25 = 0.3600 and 9/24 = 0.3750. To four decimal places, these numbers are not equivalent.

When you analyze data, it is important to be aware of **sampling errors** and nonsampling errors. The actual process of sampling causes sampling errors. For example, the sample may not be large enough. Factors not related to the sampling process cause **nonsampling errors**. A defective counting device can cause a nonsampling error.

In reality, a sample will never be exactly representative of the population so there will always be some sampling error. As a rule, the larger the sample, the smaller the sampling error.

In statistics, **a sampling bias** is created when a sample is collected from a population and some members of the population are not as likely to be chosen as others (remember, each member of the population should have an equally likely chance of being chosen). When a sampling bias happens, there can be incorrect conclusions drawn about the population that is being studied.

Example:
Exercise:

Problem:

A study is done to determine the average tuition that San Jose State undergraduate students pay per semester. Each student in the following samples is asked how much tuition he or she paid for the Fall semester. What is the type of sampling in each case?

a. A sample of 100 undergraduate San Jose State students is taken by organizing the students' names by classification (freshman, sophomore, junior, or senior), and then selecting 25 students from each.
b. A random number generator is used to select a student from the alphabetical listing of all undergraduate students in the Fall semester. Starting with that student, every 50th student is chosen until 75 students are included in the sample.
c. A completely random method is used to select 75 students. Each undergraduate student in the fall semester has the same probability of being chosen at any stage of the sampling process.
d. The freshman, sophomore, junior, and senior years are numbered one, two, three, and four, respectively. A random number generator is used to pick two of those years. All students in those two years are in the sample.
e. An administrative assistant is asked to stand in front of the library one Wednesday and to ask the first 100 undergraduate students he encounters what they paid for tuition the Fall semester. Those 100 students are the sample.

Solution:

a. stratified; b. systematic; c. simple random; d. cluster; e. convenience

Note:
Try It

You are going to use the random number generator to generate different types of samples from the data.
This table displays six sets of quiz scores (each quiz counts 10 points) for an elementary statistics class.

#1	**#2**	**#3**	**#4**	**#5**	**#6**
5	7	10	9	8	3
10	5	9	8	7	6
9	10	8	6	7	9
9	10	10	9	8	9
7	8	9	5	7	4
9	9	9	10	8	7
7	7	10	9	8	8
8	8	9	10	8	8
9	7	8	7	7	8
8	8	10	9	8	7

Instructions: Use the Random Number Generator to pick samples.
Exercise:

Problem:

1. Create a stratified sample by column. Pick three quiz scores randomly from each column.

  - Number each row one through ten.
  - On your calculator, press Math and arrow over to PRB.
  - For column 1, Press 5:randInt(and enter 1,10). Press ENTER. Record the number. Press ENTER 2 more times (even the repeats). Record these numbers. Record the three quiz scores in column one that correspond to these three numbers.
  - Repeat for columns two through six.
  - These 18 quiz scores are a stratified sample.

2. Create a cluster sample by picking two of the columns. Use the column numbers: one through six.

   - Press MATH and arrow over to PRB.
   - Press 5:randInt(and enter 1,6). Press ENTER. Record the number. Press ENTER and record that number.
   - The two numbers are for two of the columns.
   - The quiz scores (20 of them) in these 2 columns are the cluster sample.

3. Create a simple random sample of 15 quiz scores.

   - Use the numbering one through 60.
   - Press MATH. Arrow over to PRB. Press 5:randInt(and enter 1, 60).
   - Press ENTER 15 times and record the numbers.
   - Record the quiz scores that correspond to these numbers.
   - These 15 quiz scores are the systematic sample.

4. Create a systematic sample of 12 quiz scores.

   - Use the numbering one through 60.
   - Press MATH. Arrow over to PRB. Press 5:randInt(and enter 1, 60).
   - Press ENTER. Record the number and the first quiz score. From that number, count ten quiz scores and record that quiz score. Keep counting ten quiz scores and recording the quiz score until you have a sample of 12 quiz scores. You may wrap around (go back to the beginning).

Example:
Exercise:

Problem:

Determine the type of sampling used (simple random, stratified, systematic, cluster, or convenience).

a. A soccer coach selects six players from a group of boys aged eight to ten, seven players from a group of boys aged 11 to 12, and three players from a group of boys aged 13 to 14 to form a recreational soccer team.
b. A pollster interviews all human resource personnel in five different high tech companies.
c. A high school educational researcher interviews 50 high school female teachers and 50 high school male teachers.
d. A medical researcher interviews every third cancer patient from a list of cancer patients at a local hospital.
e. A high school counselor uses a computer to generate 50 random numbers and then picks students whose names correspond to the numbers.
f. A student interviews classmates in his algebra class to determine how many pairs of jeans a student owns, on the average.

Solution:

a. stratified; b. cluster; c. stratified; d. systematic; e. simple random; f.convenience

Note:
Try It
Exercise:

Problem:

Determine the type of sampling used (simple random, stratified, systematic, cluster, or convenience).

A high school principal polls 50 freshmen, 50 sophomores, 50 juniors, and 50 seniors regarding policy changes for after school activities.

Solution:

stratified

If we were to examine two samples representing the same population, even if we used random sampling methods for the samples, they would not be exactly the same. Just as there is variation in data, there is variation in samples. As you become accustomed to sampling, the variability will begin to seem natural.

Example:
Suppose ABC College has 10,000 part-time students (the population). We are interested in the average amount of money a part-time student spends on books in the fall term. Asking all 10,000 students is an almost impossible task.
Suppose we take two different samples.
First, we use convenience sampling and survey ten students from a first term organic chemistry class. Many of these students are taking first term calculus in addition to the organic chemistry class. The amount of money they spend on books is as follows:
\$128 \$87 \$173 \$116 \$130 \$204 \$147 \$189 \$93 \$153
The second sample is taken using a list of senior citizens who take P.E. classes and taking every fifth senior citizen on the list, for a total of ten senior citizens. They spend:
\$50 \$40 \$36 \$15 \$50 \$100 \$40 \$53 \$22 \$22
It is unlikely that any student is in both samples.
Exercise:

Problem:

a. Do you think that either of these samples is representative of (or is characteristic of) the entire 10,000 part-time student population?

Solution:

a. No. The first sample probably consists of science-oriented students. Besides the chemistry course, some of them are also taking first-term calculus. Books for these classes tend to be expensive. Most of these students are, more than likely, paying more than the average part-time student for their books. The second sample is a group of senior citizens who are, more than likely, taking courses for health and interest. The amount of money they spend on books is probably much less than the average parttime student. Both samples are biased. Also, in both cases, not all students have a chance to be in either sample.

Exercise:

Problem:

b. Since these samples are not representative of the entire population, is it wise to use the results to describe the entire population?

Solution:

b. No. For these samples, each member of the population did not have an equally likely chance of being chosen.

Now, suppose we take a third sample. We choose ten different part-time students from the disciplines of chemistry, math, English, psychology, sociology, history, nursing, physical education, art, and early childhood development. (We assume that these are the only disciplines in which part-time students at ABC College are enrolled and that an equal number of part-time students are enrolled in each of the disciplines.) Each student is chosen using simple random sampling. Using a calculator, random numbers are generated and a student from a particular discipline is selected if he or she has a corresponding number. The students spend the following amounts:
$180 $50 $150 $85 $260 $75 $180 $200 $200 $150

Exercise:

Problem: c. Is the sample biased?

Solution:

c. The sample is unbiased, but a larger sample would be recommended to increase the likelihood that the sample will be close to representative of the population. However, for a biased sampling technique, even a large sample runs the risk of not being representative of the population.

Students often ask if it is "good enough" to take a sample, instead of surveying the entire population. If the survey is done well, the answer is yes.

Note:
Try It
Exercise:

Problem:

A local radio station has a fan base of 20,000 listeners. The station wants to know if its audience would prefer more music or more talk shows. Asking all 20,000 listeners is an almost impossible task.

The station uses convenience sampling and surveys the first 200 people they meet at one of the station’s music concert events. 24 people said they’d prefer more talk shows, and 176 people said they’d prefer more music.

Do you think that this sample is representative of (or is characteristic of) the entire 20,000 listener population?

Solution:

The sample probably consists more of people who prefer music because it is a concert event. Also, the sample represents only those who showed up to the event earlier than the majority. The sample probably doesn’t represent the entire fan base and is probably biased towards people who would prefer music.

Note:
Collaborative Exercise

As a class, determine whether or not the following samples are representative. If they are not, discuss the reasons.

1. To find the average GPA of all students in a university, use all honor students at the university as the sample.
2. To find out the most popular cereal among young people under the age of ten, stand outside a large supermarket for three hours and speak to every twentieth child under age ten who enters the supermarket.
3. To find the average annual income of all adults in the United States, sample U.S. congressmen. Create a cluster sample by considering each state as a stratum (group). By using simple random sampling, select states to be part of the cluster. Then survey every U.S. congressman in the cluster.
4. To determine the proportion of people taking public transportation to work, survey 20 people in New York City. Conduct the survey by sitting in Central Park on a bench and interviewing every person who sits next to you.
5. To determine the average cost of a two-day stay in a hospital in Massachusetts, survey 100 hospitals across the state using simple random sampling.

Variation in Data

Variation is present in any set of data. For example, 16-ounce cans of beverage may contain more or less than 16 ounces of liquid. In one study, eight 16 ounce cans were measured and produced the following amount (in ounces) of beverage:

15.8 16.1 15.2 14.8 15.8 15.9 16.0 15.5

Measurements of the amount of beverage in a 16-ounce can may vary because different people make the measurements or because the exact amount, 16 ounces of liquid, was not put into the cans. Manufacturers regularly run tests to determine if the amount of beverage in a 16-ounce can falls within the desired range.

Be aware that as you take data, your data may vary somewhat from the data someone else is taking for the same purpose. This is completely natural. However, if two or more of you are taking the same data and get very different results, it is time for you and the others to reevaluate your data-taking methods and your accuracy.

Variation in Samples

It was mentioned previously that two or more **samples** from the same **population**, taken randomly, and having close to the same characteristics of the population will likely be different from each other. Suppose Doreen and Jung both decide to study the average amount of time students at their college sleep each night. Doreen and Jung each take samples of 500 students. Doreen uses systematic sampling and Jung uses cluster sampling. Doreen's sample will be different from Jung's sample. Even if Doreen and Jung used the same sampling method, in all likelihood their samples would be different. Neither would be wrong, however.

Think about what contributes to making Doreen’s and Jung’s samples different.

If Doreen and Jung took larger samples (i.e. the number of data values is increased), their sample results (the average amount of time a student sleeps) might be closer to the actual population average. But still, their samples would be, in all likelihood, different from each other. This **variability in samples** cannot be stressed enough.

Size of a Sample

The size of a sample (often called the number of observations) is important. The examples you have seen in this book so far have been small. Samples of only a few hundred observations, or even smaller, are sufficient for many purposes. In polling, samples that are from 1,200 to 1,500 observations are considered large enough and good enough if the survey is random and is well done. You will learn why when you study confidence intervals.

Be aware that many large samples are biased. For example, call-in surveys are invariably biased, because people choose to respond or not.

Note:

Collaborative Exercise

Divide into groups of two, three, or four. Your instructor will give each group one six-sided die. Try this experiment twice. Roll one fair die (six-sided) 20 times. Record the number of ones, twos, threes, fours, fives, and sixes you get in the following tables ("frequency" is the number of times a particular face of the die occurs):

Face on Die	**Frequency**
1	
2	
3	
4	
5	
6	

First Experiment (20 rolls)

Face on Die	**Frequency**
1	
2	
3	
4	
5	
6	

Second Experiment (20 rolls)

Did the two experiments have the same results? Probably not. If you did the experiment a third time, do you expect the results to be identical to the first or second experiment? Why or why not?

Which experiment had the correct results? They both did. The job of the statistician is to see through the variability and draw appropriate conclusions.

Critical Evaluation

We need to evaluate the statistical studies we read about critically and analyze them before accepting the results of the studies. Common problems to be aware of include

- Problems with samples: A sample must be representative of the population. A sample that is not representative of the population is biased. Biased samples that are not representative of the population give results that are inaccurate and not valid.
- Self-selected samples: Responses only by people who choose to respond, such as call-in surveys, are often unreliable.

- Sample size issues: Samples that are too small may be unreliable. Larger samples are better, if possible. In some situations, having small samples is unavoidable and can still be used to draw conclusions. Examples: crash testing cars or medical testing for rare conditions
- Undue influence: collecting data or asking questions in a way that influences the response
- Non-response or refusal of subject to participate: The collected responses may no longer be representative of the population. Often, people with strong positive or negative opinions may answer surveys, which can affect the results.
- Causality: A relationship between two variables does not mean that one causes the other to occur. They may be related (correlated) because of their relationship through a different variable.
- Self-funded or self-interest studies: A study performed by a person or organization in order to support their claim. Is the study impartial? Read the study carefully to evaluate the work. Do not automatically assume that the study is good, but do not automatically assume the study is bad either. Evaluate it on its merits and the work done.
- Misleading use of data: improperly displayed graphs, incomplete data, or lack of context
- Confounding: When the effects of multiple factors on a response cannot be separated. Confounding makes it difficult or impossible to draw valid conclusions about the effect of each factor.

References

Gallup-Healthways Well-Being Index. http://www.well-beingindex.com/default.asp (accessed May 1, 2013).

Gallup-Healthways Well-Being Index. http://www.well-beingindex.com/methodology.asp (accessed May 1, 2013).

Gallup-Healthways Well-Being Index. http://www.gallup.com/poll/146822/gallup-healthways-index-questions.aspx (accessed May 1, 2013).

Data from http://www.bookofodds.com/Relationships-Society/Articles/A0374-How-George-Gallup-Picked-the-President

Dominic Lusinchi, “’President’ Landon and the 1936 *Literary Digest* Poll: Were Automobile and Telephone Owners to Blame?” Social Science History 36, no. 1: 23-54 (2012), http://ssh.dukejournals.org/content/36/1/23.abstract (accessed May 1, 2013).

“The Literary Digest Poll,” Virtual Laboratories in Probability and Statistics http://www.math.uah.edu/stat/data/LiteraryDigest.html (accessed May 1, 2013).

“Gallup Presidential Election Trial-Heat Trends, 1936–2008,” Gallup Politics http://www.gallup.com/poll/110548/gallup-presidential-election-trialheat-trends-19362004.aspx#4 (accessed May 1, 2013).

The Data and Story Library, http://lib.stat.cmu.edu/DASL/Datafiles/USCrime.html (accessed May 1, 2013).

LBCC Distance Learning (DL) program data in 2010-2011, http://de.lbcc.edu/reports/2010-11/future/highlights.html#focus (accessed May 1, 2013).

Data from San Jose Mercury News

Section Review

Data are individual items of information that come from a population or sample. Data may be classified as qualitative, quantitative continuous, or quantitative discrete.

Because it is not practical to measure the entire population in a study, researchers use samples to represent the population. A random sample is a representative group from the population chosen by using a method that gives each individual in the population an equal chance of being included in the sample. Random sampling methods include simple random sampling, stratified sampling, cluster sampling, and systematic sampling. Convenience sampling is a nonrandom method of choosing a sample that often produces biased data.

Samples that contain different individuals result in different data. This is true even when the samples are well-chosen and representative of the population. When properly selected, larger samples model the population more closely than smaller samples. There are many different potential problems that can affect the reliability of a sample. Statistical data needs to be critically analyzed, not simply accepted.

Practice

Exercise:

Problem: “Number of times per week” is what type of data?

a. qualitative b. quantitative discrete c. quantitative continuous

Use the following information to answer the next four exercises: A study was done to determine the age, number of times per week, and the duration (amount of time) of residents using a local park in San Antonio, Texas. The first house in the neighborhood around the park was selected randomly, and then the resident of every eighth house in the neighborhood around the park was interviewed.

Exercise:

Problem: The sampling method was

a. simple random b. systematic c. stratified d. cluster

Solution:

b

Exercise:

Problem: “Duration (amount of time)” is what type of data?

a. qualitative b. quantitative discrete c. quantitative continuous

Exercise:

Problem:

The colors of the houses around the park are what kind of data?

a. qualitative b. quantitative discrete c. quantitative continuous

---

Solution:

a

Exercise:

Problem: The population is _______________________

Exercise:

Problem:

The following table contains the total number of deaths worldwide as a result of earthquakes from 2000 to 2012.

Year	**Total Number of Deaths**
2000	231
2001	21,357
2002	11,685
2003	33,819
2004	228,802
2005	88,003

Year	**Total Number of Deaths**
2006	6,605
2007	712
2008	88,011
2009	1,790
2010	320,120
2011	21,953
2012	768
Total	**823,856**

Use [link] to answer the following questions.

a. What is the proportion of deaths between 2007 and 2012?
b. What percent of deaths occurred before 2001?
c. What is the percent of deaths that occurred in 2003 or after 2010?
d. What is the fraction of deaths that happened before 2012?
e. What kind of data is the number of deaths?
f. Earthquakes are quantified according to the amount of energy they produce (examples are 2.1, 5.0, 6.7). What type of data is that?
g. What contributed to the large number of deaths in 2010? In 2004? Explain.

Solution:

a. 0.526
b. 0.03%
c. 6.86%
d. $\frac{823{,}088}{823{,}856}$

e. quantitative discrete
f. quantitative continuous
g. In both years, underwater earthquakes produced massive tsunamis.

For the following four exercises, determine the type of sampling used (simple random, stratified, systematic, cluster, or convenience).

Exercise:

Problem:

A group of test subjects is divided into twelve groups; then four of the groups are chosen at random.

Exercise:

Problem:

A market researcher polls every tenth person who walks into a store.

Solution:

systematic

Exercise:

Problem:

The first 50 people who walk into a sporting event are polled on their television preferences.

Exercise:

Problem:

A computer generates 100 random numbers, and 100 people whose names correspond with the numbers on the list are chosen.

Solution:

simple random

Use the following information to answer the next seven exercises: Studies are often done by pharmaceutical companies to determine the effectiveness of a treatment program. Suppose that a new AIDS antibody drug is currently under study. It is given to patients once the AIDS symptoms have revealed themselves. Of interest is the average (mean) length of time in months patients live once starting the treatment. Two researchers each follow a different set of 40 AIDS patients from the start of treatment until their deaths. The following data (in months) are collected.

Researcher A: 3; 4; 11; 15; 16; 17; 22; 44; 37; 16; 14; 24; 25; 15; 26; 27; 33; 29; 35; 44; 13; 21; 22; 10; 12; 8; 40; 32; 26; 27; 31; 34; 29; 17; 8; 24; 18; 47; 33; 34

Researcher B: 3; 14; 11; 5; 16; 17; 28; 41; 31; 18; 14; 14; 26; 25; 21; 22; 31; 2; 35; 44; 23; 21; 21; 16; 12; 18; 41; 22; 16; 25; 33; 34; 29; 13; 18; 24; 23; 42; 33; 29
Exercise:

Problem: Complete the tables using the data provided:

Survival Length (in months)	Frequency	Relative Frequency	Cumulative Relative Frequency
0.5–6.5			
6.5–12.5			
12.5–18.5			
18.5–24.5			

Survival Length (in months)	Frequency	Relative Frequency	Cumulative Relative Frequency
24.5–30.5			
30.5–36.5			
36.5–42.5			
42.5–48.5			

Researcher A

Survival Length (in months)	Frequency	Relative Frequency	Cumulative Relative Frequency
0.5–6.5			
6.5–12.5			
12.5–18.5			
18.5–24.5			
24.5–30.5			
30.5–36.5			
36.5-45.5			

Researcher B

Exercise:

Problem:

Determine what the key term data refers to in the above example for Researcher A.

Solution:

values for X, such as 3, 4, 11, and so on

Exercise:

Problem: List two reasons why the data may differ.

Exercise:

Problem:

Can you tell if one researcher is correct and the other one is incorrect? Why?

Solution:

No, we do not have enough information to make such a claim.

Exercise:

Problem: Would you expect the data to be identical? Why or why not?

Exercise:

Problem: How might the researchers gather random data?

Solution:

Take a simple random sample from each group. One way is by assigning a number to each patient and using a random number generator to randomly select patients.

Exercise:

Problem:

Suppose that the first researcher conducted his survey by randomly choosing one state in the nation and then randomly picking 40 patients from that state. What sampling method would that researcher have used?

Exercise:

Problem:

Suppose that the second researcher conducted his survey by choosing 40 patients he knew. What sampling method would that researcher have used? What concerns would you have about this data set, based upon the data collection method?

Solution:

This would be convenience sampling and is not random.

Use the following data to answer the next five exercises: Two researchers are gathering data on hours of video games played by school-aged children and young adults. They each randomly sample different groups of 150 students from the same school. They collect the following data.

Hours Played per Week	Frequency	Relative Frequency	Cumulative Relative Frequency
0–2	26	0.17	0.17
2–4	30	0.20	0.37
4–6	49	0.33	0.70

Hours Played per Week	Frequency	Relative Frequency	Cumulative Relative Frequency
6–8	25	0.17	0.87
8–10	12	0.08	0.95
10–12	8	0.05	1

Researcher A

Hours Played per Week	Frequency	Relative Frequency	Cumulative Relative Frequency
0–2	48	0.32	0.32
2–4	51	0.34	0.66
4–6	24	0.16	0.82
6–8	12	0.08	0.90
8–10	11	0.07	0.97
10–12	4	0.03	1

Researcher B

Exercise:

Problem: Give a reason why the data may differ.

Exercise:

Problem:

Would the sample size be large enough if the population is the students in the school?

Solution:

Yes, the sample size of 150 would be large enough to reflect a population of one school.

Exercise:

Problem:

Would the sample size be large enough if the population is school-aged children and young adults in the United States?

Exercise:

Problem:

Researcher A concludes that most students play video games between four and six hours each week. Researcher B concludes that most students play video games between two and four hours each week. Who is correct?

Solution:

Even though the specific data support each researcher’s conclusions, the different results suggest that more data need to be collected before the researchers can reach a conclusion.

Exercise:

Problem:

As part of a way to reward students for participating in the survey, the researchers gave each student a gift card to a video game store. Would this affect the data if students knew about the award before the study?

Use the following data to answer the next five exercises: A pair of studies was performed to measure the effectiveness of a new software program designed to help stroke patients regain their problem-solving skills. Patients were asked to use the software program twice a day, once in the morning and once in the evening. The studies observed 200 stroke patients recovering over a period of several weeks. The first study collected the data in [link]. The second study collected the data in [link].

Group	**Showed improvement**	**No improvement**	**Deterioration**
Used program	142	43	15
Did not use program	72	110	18

Group	**Showed improvement**	**No improvement**	**Deterioration**
Used program	105	74	19
Did not use program	89	99	12

Exercise:

Problem: Given what you know, which study is correct?

Solution:

There is not enough information given to judge if either one is correct or incorrect.

Exercise:

Problem:

The first study was performed by the company that designed the software program. The second study was performed by the American Medical Association. Which study is more reliable?

Exercise:

Problem:

Both groups that performed the study concluded that the software works. Is this accurate?

Solution:

The software program seems to work because the second study shows that more patients improve while using the software than not. Even though the difference is not as large as that in the first study, the results from the second study are likely more reliable and still show improvement.

Exercise:

Problem:

The company takes the two studies as proof that their software causes mental improvement in stroke patients. Is this a fair statement?

Exercise:

Problem:

Patients who used the software were also a part of an exercise program whereas patients who did not use the software were not. Does this change the validity of the conclusions from [link]?

Solution:

Yes, because we cannot tell if the improvement was due to the software or the exercise; the data is confounded, and a reliable conclusion cannot be drawn. New studies should be performed.

Exercise:

Problem:

Is a sample size of 1,000 a reliable measure for a population of 5,000?

Exercise:

Problem:

Is a sample of 500 volunteers a reliable measure for a population of 2,500?

Solution:

No, even though the sample is large enough, the fact that the sample consists of volunteers makes it a self-selected sample, which is not reliable.

Exercise:

Problem:

A question on a survey reads: "Do you prefer the delicious taste of Brand X or the taste of Brand Y?" Is this a fair question?

Exercise:

Problem: Is a sample size of two representative of a population of five?

Solution:

No, even though the sample is a large portion of the population, two responses are not enough to justify any conclusions. Because the population is so small, it would be better to include everyone in the population to get the most accurate data.

Exercise:

Problem:

Is it possible for two experiments to be well run with similar sample sizes to get different data?

HOMEWORK

For the following exercises, identify the type of data that would be used to describe a response (quantitative discrete, quantitative continuous, or qualitative), and give an example of the data.

Exercise:

Problem: number of tickets sold to a concert

Solution:

quantitative discrete, 150

Exercise:

Problem: percent of body fat

Exercise:

Problem: favorite baseball team

Solution:

qualitative, Oakland A’s

Exercise:

Problem: time in line to buy groceries

Exercise:

Problem: number of students enrolled at Evergreen Valley College

---

Solution:

quantitative discrete, 11,234 students

Exercise:

Problem: most-watched television show

Exercise:

Problem: brand of toothpaste

Solution:

qualitative, Crest

Exercise:

Problem: distance to the closest movie theatre

Exercise:

Problem: age of executives in Fortune 500 companies

Solution:

quantitative continuous, 47.3 years

Exercise:

Problem: number of competing computer spreadsheet software packages

Use the following information to answer the next two exercises: A study was done to determine the age, number of times per week, and the duration (amount of time) of resident use of a local park in San Jose. The first house in the neighborhood around the park was selected randomly and then every 8th house in the neighborhood around the park was interviewed.

Exercise:

Problem: “Number of times per week” is what type of data?

a. qualitative
b. quantitative discrete
c. quantitative continuous

Solution:

b

Exercise:

Problem: “Duration (amount of time)” is what type of data?

a. qualitative
b. quantitative discrete
c. quantitative continuous

Exercise:

Problem:

Airline companies are interested in the consistency of the number of babies on each flight, so that they have adequate safety equipment. Suppose an airline conducts a survey. Over Thanksgiving weekend, it surveys six flights from Boston to Salt Lake City to determine the number of babies on the flights. It determines the amount of safety equipment needed by the result of that study.

a. Using complete sentences, list three things wrong with the way the survey was conducted.

b. Using complete sentences, list three ways that you would improve the survey if it were to be repeated.

Solution:

a. The survey was conducted using six similar flights.
   The survey would not be a true representation of the entire population of air travelers.
   Conducting the survey on a holiday weekend will not produce representative results.
b. Conduct the survey during different times of the year.
   Conduct the survey using flights to and from various locations.
   Conduct the survey on different days of the week.

Exercise:

Problem:

Suppose you want to determine the mean number of students per statistics class in your state. Describe a possible sampling method in three to five complete sentences. Make the description detailed.

Exercise:

Problem:

Suppose you want to determine the mean number of cans of soda drunk each month by students in their twenties at your school. Describe a possible sampling method in three to five complete sentences. Make the description detailed.

Solution:

Answers will vary. Sample Answer: You could use a systematic sampling method. Stop the tenth person as they leave one of the buildings on campus at 9:50 in the morning. Then stop the tenth person as they leave a different building on campus at 1:50 in the afternoon.

Exercise:

Problem:

List some practical difficulties involved in getting accurate results from a telephone survey.

Exercise:

Problem:

List some practical difficulties involved in getting accurate results from a mailed survey.

Solution:

Answers will vary. Sample Answer: Many people will not respond to mail surveys. If they do respond to the surveys, you can’t be sure who is responding. In addition, mailing lists can be incomplete.

Exercise:

Problem:

With your classmates, brainstorm some ways you could overcome these problems if you needed to conduct a phone or mail survey.

Exercise:

Problem:

The instructor takes her sample by gathering data on five randomly selected students from each Lake Tahoe Community College math class. The type of sampling she used is

a. cluster sampling
b. stratified sampling
c. simple random sampling
d. convenience sampling

Solution:

b

Exercise:

Problem:

A study was done to determine the age, number of times per week, and the duration (amount of time) of residents using a local park in San Jose. The first house in the neighborhood around the park was selected randomly and then every eighth house in the neighborhood around the park was interviewed. The sampling method was:

a. simple random
b. systematic
c. stratified
d. cluster

Exercise:

Problem:

Name the sampling method used in each of the following situations:

a. A woman in the airport is handing out questionnaires to travelers asking them to evaluate the airport's service. She does not ask travelers who are hurrying through the airport with their hands full of luggage, but instead asks all travelers who are sitting near gates and not taking naps while they wait.
b. A teacher wants to know if her students are doing homework, so she randomly selects rows two and five and then calls on all students in row two and all students in row five to present the solutions to homework problems to the class.
c. The marketing manager for an electronics chain store wants information about the ages of its customers. Over the next two weeks, at each store location, 100 randomly selected customers are given questionnaires to fill out asking for information about age, as well as about other variables of interest.
d. The librarian at a public library wants to determine what proportion of the library users are children. The librarian has a tally sheet on which she marks whether books are checked out by an adult or a child. She records this data for every fourth patron who checks out books.

e. A political party wants to know the reaction of voters to a debate between the candidates. The day after the debate, the party's polling staff calls 1,200 randomly selected phone numbers. If a registered voter answers the phone or is available to come to the phone, that registered voter is asked whom he or she intends to vote for and whether the debate changed his or her opinion of the candidates.

Solution:

convenience cluster stratified systematic simple random

Exercise:

Problem:

A "random survey" was conducted of 3,274 people of the "microprocessor generation" (people born since 1971, the year the microprocessor was invented). It was reported that 48% of those individuals surveyed stated that if they had $2,000 to spend, they would use it for computer equipment. Also, 66% of those surveyed considered themselves relatively savvy computer users.

a. Do you consider the sample size large enough for a study of this type? Why or why not?
b. Based on your "gut feeling," do you believe the percents accurately reflect the U.S. population for those individuals born since 1971? If not, do you think the percents of the population are actually higher or lower than the sample statistics? Why?
Additional information: The survey, reported by Intel Corporation, was filled out by individuals who visited the Los Angeles Convention Center to see the Smithsonian Institute's road show called "America's Smithsonian."
c. With this additional information, do you feel that all demographic and ethnic groups were equally represented at the event? Why or why not?
d. With the additional information, comment on how accurately you think the sample statistics reflect the population parameters.

Exercise:

Problem:

The Gallup-Healthways Well-Being Index is a survey that follows trends of U.S. residents on a regular basis. There are six areas of health and wellness covered in the survey: Life Evaluation, Emotional Health, Physical Health, Healthy Behavior, Work Environment, and Basic Access. Some of the questions used to measure the Index are listed below.

Identify the type of data obtained from each question used in this survey: qualitative, quantitative discrete, or quantitative continuous.

a. Do you have any health problems that prevent you from doing any of the things people your age can normally do?
b. During the past 30 days, for about how many days did poor health keep you from doing your usual activities?
c. In the last seven days, on how many days did you exercise for 30 minutes or more?
d. Do you have health insurance coverage?

Solution:

a. qualitative
b. quantitative discrete
c. quantitative discrete
d. qualitative

Exercise:

Problem:

In advance of the 1936 Presidential Election, a magazine titled Literary Digest released the results of an opinion poll predicting that the republican candidate Alf Landon would win by a large margin. The magazine sent post cards to approximately 10,000,000 prospective voters. These prospective voters were selected from the subscription list of the magazine, from automobile registration lists, from phone lists, and from club membership lists. Approximately 2,300,000 people returned the postcards.

a. Think about the state of the United States in 1936. Explain why a sample chosen from magazine subscription lists, automobile registration lists, phone books, and club membership lists was not representative of the population of the United States at that time.
b. What effect does the low response rate have on the reliability of the sample?
c. Are these problems examples of sampling error or nonsampling error?
d. During the same year, George Gallup conducted his own poll of 30,000 prospective voters. His researchers used a method they called "quota sampling" to obtain survey answers from specific subsets of the population. Quota sampling is an example of which sampling method described in this module?

Exercise:

Problem:

A scholarly article about response rates begins with the following quote:

“Declining contact and cooperation rates in random digit dial (RDD) national telephone surveys raise serious concerns about the validity of estimates drawn from such research.”[footnote]

Scott Keeter et al., “Gauging the Impact of Growing Nonresponse on Estimates from a National RDD Telephone Survey,” Public Opinion Quarterly 70 no. 5 (2006), http://poq.oxfordjournals.org/content/70/5/759.full (accessed May 1, 2013).

The Pew Research Center for People and the Press admits:

“The percentage of people we interview – out of all we try to interview – has been declining over the past decade or more.”[footnote]
Frequently Asked Questions, Pew Research Center for the People & the Press, http://www.people-press.org/methodology/frequently-asked-questions/#dont-you-have-trouble-getting-people-to-answer-your-polls (accessed May 1, 2013).

a. What are some reasons for the decline in response rate over the past decade?
b. Explain why researchers are concerned with the impact of the declining response rate on public opinion polls.

Solution:

a. Possible reasons: increased use of caller id, decreased use of landlines, increased use of private numbers, voice mail, privacy managers, hectic nature of personal schedules, decreased willingness to be interviewed
b. When a large number of people refuse to participate, then the sample may not have the same characteristics of the population. Perhaps the majority of people willing to participate are doing so because they feel strongly about the subject of the survey.

Bringing It Together

Exercise:

Problem:

Seven hundred and seventy-one distance learning students at Long Beach City College responded to surveys in the 2010-11 academic year. Highlights of the summary report are listed in the following table.

Have computer at home	96%
Unable to come to campus for classes	65%
Age 41 or over	24%
Would like LBCC to offer more DL courses	95%
Took DL classes due to a disability	17%
Live at least 16 miles from campus	13%
Took DL courses to fulfill transfer requirements	71%

LBCC Distance Learning Survey Results

a. What percent of the students surveyed do not have a computer at home?
b. About how many students in the survey live at least 16 miles from campus?
c. If the same survey were done at Great Basin College in Elko, Nevada, do you think the percentages would be the same? Why?

Exercise:

Problem:

Several online textbook retailers advertise that they have lower prices than on-campus bookstores. However, an important factor is whether the Internet retailers actually have the textbooks that students need in stock. Students need to be able to get textbooks promptly at the beginning of the college term. If the book is not available, then a student would not be able to get the textbook at all, or might get a delayed delivery if the book is back ordered.

A college newspaper reporter is investigating textbook availability at online retailers. He decides to investigate one textbook for each of the following seven subjects: calculus, biology, chemistry, physics, statistics, geology, and general engineering. He consults textbook industry sales data and selects the most popular nationally used textbook in each of these subjects. He visits websites for a random sample of major online textbook sellers and looks up each of these seven textbooks to see if they are available in stock for quick delivery through these retailers. Based on his investigation, he writes an article in which he draws conclusions about the overall availability of all college textbooks through online textbook retailers.

Write an analysis of his study that addresses the following issues: Is his sample representative of the population of all college textbooks? Explain why or why not. Describe some possible sources of bias in this study, and how it might affect the results of the study. Give some suggestions about what could be done to improve the study.

Solution:

Answers will vary. Sample answer: The sample is not representative of the population of all college textbooks. Two reasons why it is not representative are that he only sampled seven subjects and he only investigated one textbook in each subject. There are several possible sources of bias in the study. The seven subjects that he investigated are all in mathematics and the sciences; there are many subjects in the humanities, social sciences, and other subject areas, (for example: literature, art, history, psychology, sociology, business) that he did not investigate at all. It may be that different subject areas exhibit different patterns of textbook availability, but his sample would not detect such results.

He also looked only at the most popular textbook in each of the subjects he investigated. The availability of the most popular textbooks may differ from the availability of other textbooks in one of two ways:

- the most popular textbooks may be more readily available online, because more new copies are printed, and more students nationwide are selling back their used copies OR
- the most popular textbooks may be harder to find available online, because more student demand exhausts the supply more quickly.

In reality, many college students do not use the most popular textbook in their subject, and this study gives no useful information about the situation for those less popular textbooks.

He could improve this study by:

- expanding the selection of subjects he investigates so that it is more representative of all subjects studied by college students, and
- expanding the selection of textbooks he investigates within each subject to include a mixed representation of both the most popular and less popular textbooks.

Glossary

Cluster Sampling

a method for selecting a random sample and dividing the population into groups (clusters); use simple random sampling to select a set of clusters. Every individual in the chosen clusters is included in the sample.

Continuous Random Variable

a random variable (RV) whose outcomes are measured; the height of trees in the forest is a continuous RV.

Convenience Sampling

a nonrandom method of selecting a sample; this method selects individuals that are easily accessible and may result in biased data.

Discrete Random Variable

a random variable (RV) whose outcomes are counted

Nonsampling Error

an issue that affects the reliability of sampling data other than natural variation; it includes a variety of human errors including poor study design, biased sampling methods, inaccurate information provided by study participants, data entry errors, and poor analysis.

Qualitative Data

Data in which each data value falls into a particular category. Often referred to as categorical data.

Quantitative Data
Data that consists of numeric values, which are the result of measuring or counting. Often referred to as numerical data.

Random Sampling
a method of selecting a sample that gives every member of the population an equal chance of being selected.

Sampling Bias
not all members of the population are equally likely to be selected

Sampling Error
the natural variation that results from selecting a sample to represent a larger population; this variation decreases as the sample size increases, so selecting larger samples reduces sampling error.

Sampling with Replacement
Once a member of the population is selected for inclusion in a sample, that member is returned to the population for the selection of the next individual.

Sampling without Replacement
A member of the population may be chosen for inclusion in a sample only once. If chosen, the member is not returned to the population before the next selection.

Simple Random Sampling
a straightforward method for selecting a random sample; give each member of the population a number. Use a random number generator to select a set of labels. These randomly selected labels identify the members of your sample.

Stratified Sampling
a method for selecting a random sample used to ensure that subgroups of the population are represented adequately; divide the population into

groups (strata). Use simple random sampling to identify a proportionate number of individuals from each stratum.

Systematic Sampling

a method for selecting a random sample; list the members of the population. Use simple random sampling to select a starting point in the population. Let k = (number of individuals in the population)/(number of individuals needed in the sample). Choose every kth individual in the list starting with the one that was randomly selected. If necessary, return to the beginning of the population list to complete your sample.

Frequency and Frequency Tables

Once you have a set of data, you will need to organize it so that you can analyze how frequently each datum occurs in the set. However, when calculating the frequency, you may need to round your answers so that they are as precise as possible.

Answers and Rounding Off

A simple way to round off answers is to carry your final answer one more decimal place than was present in the original data. Round off only the final answer. Do not round off any intermediate results, if possible. If it becomes necessary to round off intermediate results, carry them to at least twice as many decimal places as the final answer. For example, the average of the three quiz scores four, six, and nine is 6.3, rounded off to the nearest tenth, because the data are whole numbers. Most answers will be rounded off in this manner.

Frequency

Twenty students were asked how many hours they worked per day. Their responses, in hours, are as follows: 56332475235654435253.

The following table lists the different data values in ascending order and their frequencies.

DATA VALUE	**FREQUENCY**
2	3
3	5
4	3
5	6
6	2
7	1

Frequency Table of Student Work Hours

A **frequency** is the number of times a value of the data occurs. According to [link], there are three students who work two hours, five students who work three hours, and so on. The sum of the values in the frequency column, 20, represents the total number of students included in the sample.

A **relative frequency** is the ratio (fraction or proportion) of the number of times a value of the data occurs in the set of all outcomes to the total number of outcomes. To find the relative frequencies, divide each frequency by the total number of students in the sample–in this case, 20. Relative frequencies can be written as fractions, percents, or decimals.

DATA VALUE	**FREQUENCY**	**RELATIVE FREQUENCY**
2	3	$\frac{3}{20}$ or 0.15
3	5	$\frac{5}{20}$ or 0.25
4	3	$\frac{3}{20}$ or 0.15
5	6	$\frac{6}{20}$ or 0.30
6	2	$\frac{2}{20}$ or 0.10
7	1	$\frac{1}{20}$ or 0.05

Frequency Table of Student Work Hours with Relative Frequencies

The sum of the values in the relative frequency column of [link] is $\frac{20}{20}$, or 1.

Cumulative relative frequency is the accumulation of the previous relative frequencies. To find the cumulative relative frequencies, add all the previous relative frequencies to the relative frequency for the current row, as shown in the following table.

DATA VALUE	**FREQUENCY**	**RELATIVE FREQUENCY**	**CUMULATIVE RELATIVE FREQUENCY**
2	3	$\frac{3}{20}$ or 0.15	0.15
3	5	$\frac{5}{20}$ or 0.25	0.15 + 0.25 = 0.40

DATA VALUE	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
4	3	$\frac{3}{20}$ or 0.15	0.40 + 0.15 = 0.55
5	6	$\frac{6}{20}$ or 0.30	0.55 + 0.30 = 0.85
6	2	$\frac{2}{20}$ or 0.10	0.85 + 0.10 = 0.95
7	1	$\frac{1}{20}$ or 0.05	0.95 + 0.05 = 1.00

Frequency Table of Student Work Hours with Relative and Cumulative Relative Frequencies

The last entry of the cumulative relative frequency column is one, indicating that one hundred percent of the data has been accumulated.

Note:
NOTE
Because of rounding, the relative frequency column may not always sum to one, and the last entry in the cumulative relative frequency column may not be one. However, they each should be close to one.

The following table represents the heights, in inches, of a sample of 100 male semiprofessional soccer players.

HEIGHTS (INCHES)	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
59.95–61.95	5	$\frac{5}{100}$ = 0.05	0.05
61.95–63.95	3	$\frac{3}{100}$ = 0.03	0.05 + 0.03 = 0.08
63.95–65.95	15	$\frac{15}{100}$ = 0.15	0.08 + 0.15 = 0.23
65.95–67.95	40	$\frac{40}{100}$ = 0.40	0.23 + 0.40 = 0.63

HEIGHTS (INCHES)	**FREQUENCY**	**RELATIVE FREQUENCY**	**CUMULATIVE RELATIVE FREQUENCY**
67.95–69.95	17	$\frac{17}{100} = 0.17$	0.63 + 0.17 = 0.80
69.95–71.95	12	$\frac{12}{100} = 0.12$	0.80 + 0.12 = 0.92
71.95–73.95	7	$\frac{7}{100} = 0.07$	0.92 + 0.07 = 0.99
73.95–75.95	1	$\frac{1}{100} = 0.01$	0.99 + 0.01 = 1.00
	Total = 100	**Total = 1.00**	

Frequency Table of Soccer Player Height

The data in this table have been **grouped** into the following intervals:

- 59.95 to 61.95 inches
- 61.95 to 63.95 inches
- 63.95 to 65.95 inches
- 65.95 to 67.95 inches
- 67.95 to 69.95 inches
- 69.95 to 71.95 inches
- 71.95 to 73.95 inches
- 73.95 to 75.95 inches

In this sample, there are **five** players whose heights fall within the interval 59.95–61.95 inches, **three** players whose heights fall within the interval 61.95–63.95 inches, **15** players whose heights fall within the interval 63.95–65.95 inches, **40** players whose heights fall within the interval 65.95–67.95 inches, **17** players whose heights fall within the interval 67.95–69.95 inches, **12** players whose heights fall within the interval 69.95–71.95, **seven** players whose heights fall within the interval 71.95–73.95, and **one** player whose heights fall within the interval 73.95–75.95. All heights fall between the endpoints of an interval and not at the endpoints.

Example:
Exercise:

Problem: From [link], find the percentage of heights that are less than 65.95 inches.

Solution:

If you look at the first, second, and third rows, the heights are all less than 65.95 inches. There are 5 + 3 + 15 = 23 players whose heights are less than 65.95 inches. The percentage of heights less than 65.95 inches is then $\frac{23}{100}$ or 23%. This percentage is the cumulative relative frequency entry in the third row.

Note:
Try It
Exercise:

Problem:

The following table shows the amount, in inches, of annual rainfall in a sample of towns.

Rainfall (Inches)	**Frequency**	**Relative Frequency**	**Cumulative Relative Frequency**
2.95–4.97	6	$\frac{6}{50}$ = 0.12	0.12
4.97–6.99	7	$\frac{7}{50}$ = 0.14	0.12 + 0.14 = 0.26
6.99–9.01	15	$\frac{15}{50}$ = 0.30	0.26 + 0.30 = 0.56
9.01–11.03	8	$\frac{8}{50}$ = 0.16	0.56 + 0.16 = 0.72
11.03–13.05	9	$\frac{9}{50}$ = 0.18	0.72 + 0.18 = 0.90
13.05–15.07	5	$\frac{5}{50}$ = 0.10	0.90 + 0.10 = 1.00
	Total = 50	Total = 1.00	

From the table, find the percentage of rainfall that is less than 9.01 inches.

Solution:

0.56 or 56%

Example:
Exercise:

Problem:

From [link], find the percentage of heights that fall between 61.95 and 65.95 inches.

Solution:

Add the relative frequencies in the second and third rows: 0.03 + 0.15 = 0.18 or 18%.

Note:
Try It
Exercise:

Problem: From [link], find the percentage of rainfall that is between 6.99 and 13.05 inches.

Solution:

0.30 + 0.16 + 0.18 = 0.64 or 64%

Example:
Exercise:

Problem:

Use the heights of the 100 male semiprofessional soccer players in [link]. Fill in the blanks and check your answers.

a. The percentage of heights that are from 67.95 to 71.95 inches is: ____.
b. The percentage of heights that are from 67.95 to 73.95 inches is: ____.
c. The percentage of heights that are more than 65.95 inches is: ____.
d. The number of players in the sample who are between 61.95 and 71.95 inches tall is: ____.
e. What kind of data are the heights?
f. Describe how you could gather this data (the heights) so that the data are characteristic of all male semiprofessional soccer players.

Remember, you **count frequencies**. To find the relative frequency, divide the frequency by the total number of data values. To find the cumulative relative frequency, add all of the previous relative frequencies to the relative frequency for the current row.

Solution:

a. 29%
b. 36%
c. 77%

d. 87
e. quantitative continuous
f. get rosters from each team and choose a simple random sample from each

Note:
Try It
Exercise:

Problem:

From [link], find the number of towns that have rainfall between 2.95 and 9.01 inches.

Solution:

6 + 7 + 15 = 28 towns

Note:
Collaborative Exercise
In your class, have someone conduct a survey of the number of siblings (brothers and sisters) each student has. Create a frequency table. Add to it a relative frequency column and a cumulative relative frequency column. Answer the following questions:

1. What percentage of the students in your class have no siblings?
2. What percentage of the students have from one to three siblings?
3. What percentage of the students have fewer than three siblings?

Example:
Nineteen people were asked how many miles, to the nearest mile, they commute to work each day. The data are as follows: 2 5 7 3 2 10 18 15 20 7 10 18 5 12 13 12 4 5 10. The following table was produced:

DATA	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY

DATA	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
3	3	$\frac{3}{19}$	0.1579
4	1	$\frac{1}{19}$	0.2105
5	3	$\frac{3}{19}$	0.1579
7	2	$\frac{2}{19}$	0.2632
10	3	$\frac{4}{19}$	0.4737
12	2	$\frac{2}{19}$	0.7895
13	1	$\frac{1}{19}$	0.8421
15	1	$\frac{1}{19}$	0.8948
18	1	$\frac{1}{19}$	0.9474
20	1	$\frac{1}{19}$	1.0000

Frequency of Commuting Distances

Exercise:

Problem:

a. Is the table correct? If it is not correct, what is wrong?
b. True or False: Three percent of the people surveyed commute three miles. If the statement is not correct, what should it be? If the table is incorrect, make the corrections.
c. What fraction of the people surveyed commute five or seven miles?
d. What fraction of the people surveyed commute 12 miles or more? Less than 12 miles? Between five and 13 miles (not including five and 13 miles)?

Solution:

a. No. The frequency column sums to 18, not 19. There should be a row for 2 miles, with a frequency of 2, and the frequency for 3 miles should be 1. Not all cumulative relative frequencies are correct.
b. False. The frequency for three miles should be one, which means 1/19 or about 5% of people surveyed commute 3 miles. The cumulative relative frequency column should read: 0.1052, 0.1579, 0.2105, 0.3684, 0.4737, 0.6316, 0.7368, 0.7895, 0.8421, 0.9474, 1.0000.

c. $\frac{5}{19}$
d. $\frac{7}{19}, \frac{12}{19}, \frac{7}{19}$

Note:
Try It
Exercise:

Problem:

[link] represents the amount, in inches, of annual rainfall in a sample of towns. What fraction of towns surveyed get between 11.03 and 13.05 inches of rainfall each year?

Solution:

$\frac{9}{50}$

Example:
The following table contains the total number of deaths worldwide as a result of earthquakes for the period from 2000 to 2012.

Year	Total Number of Deaths
2000	231
2001	21,357
2002	11,685
2003	33,819
2004	228,802
2005	88,003
2006	6,605
2007	712

Year	Total Number of Deaths
2008	88,011
2009	1,790
2010	320,120
2011	21,953
2012	768
Total	823,356

Exercise:

Problem: Answer the following questions.

a. What is the frequency of deaths measured from 2006 through 2009?
b. What percentage of deaths occurred after 2009?
c. What is the relative frequency of deaths that occurred in 2003 or earlier?
d. What is the percentage of deaths that occurred in 2004?
e. What kind of data are the numbers of deaths?
f. The Richter scale is used to quantify the energy produced by an earthquake. Examples of Richter scale numbers are 2.3, 4.0, 6.1, and 7.0. What kind of data are these numbers?

Solution:

a. 97,118 (11.8%)
b. 41.6%
c. 67,092/823,356 or 0.081 or 8.1 %
d. 27.8%
e. Quantitative discrete
f. Quantitative continuous

Note:
Try It
Exercise:

Problem:

The following table contains the total number of fatal motor vehicle traffic crashes in the United States for the period from 1994 to 2011.

Year	Total Number of Crashes	Year	Total Number of Crashes
1994	36,254	2004	38,444
1995	37,241	2005	39,252
1996	37,494	2006	38,648
1997	37,324	2007	37,435
1998	37,107	2008	34,172
1999	37,140	2009	30,862
2000	37,526	2010	30,296
2001	37,862	2011	29,757
2002	38,491	Total	653,782
2003	38,477		

Answer the following questions.

a. What is the frequency of deaths measured from 2000 through 2004?
b. What percentage of deaths occurred after 2006?
c. What is the relative frequency of deaths that occurred in 2000 or before?
d. What is the percentage of deaths that occurred in 2011?
e. What is the cumulative relative frequency for 2006? Explain what this number tells you about the data.

Solution:

a. 190,800 (29.2%)
b. 24.9%
c. 260,086/653,782 or 39.8%
d. 4.6%
e. 75.1% of all fatal traffic crashes for the period from 1994 to 2011 happened from 1994 to 2006.

References

"State & County QuickFacts," U.S. Census Bureau. http://quickfacts.census.gov/qfd/download_data.html (accessed May 1, 2013).

“State & County QuickFacts: Quick, easy access to facts about people, business, and geography,” U.S. Census Bureau. http://quickfacts.census.gov/qfd/index.html (accessed May 1, 2013).

“Table 5: Direct hits by mainland United States Hurricanes (1851-2004),” National Hurricane Center, http://www.nhc.noaa.gov/gifs/table5.gif (accessed May 1, 2013).

“Levels of Measurement,” http://infinity.cos.edu/faculty/woodbury/stats/tutorial/Data_Levels.htm (accessed May 1, 2013).

Courtney Taylor, “Levels of Measurement,” about.com, http://statistics.about.com/od/HelpandTutorials/a/Levels-Of-Measurement.htm (accessed May 1, 2013).

David Lane. “Levels of Measurement,” Connexions, http://cnx.org/content/m10809/latest/ (accessed May 1, 2013).

Section Review

Some calculations generate numbers that are artificially precise. It is not necessary to report a value to eight decimal places when the measures that generated that value were only accurate to the nearest tenth. Round off your final answer to one more decimal place than was present in the original data. This means that if you have data measured to the nearest tenth of a unit, report the final statistic to the nearest hundredth.

In addition to rounding your answers, you can measure your data using the following four levels of measurement.

When organizing data, it is important to know how many times a value appears. How many statistics students study five hours or more for an exam? What percent of families on our block own two pets? Frequency, relative frequency, and cumulative relative frequency are measures that answer questions like these.

HOMEWORK

Exercise:

Problem:

Fifty part-time students were asked how many courses they were taking this term. The (incomplete) results are shown below:

# of Courses	Frequency	Relative Frequency	Cumulative Relative Frequency

# of Courses	Frequency	Relative Frequency	Cumulative Relative Frequency
1	30	0.6	
2	15		
3			

Part-time Student Course Loads

a. Fill in the blanks in [link].
b. What percent of students take exactly two courses?
c. What percent of students take one or two courses?

Exercise:

Problem:

Sixty adults with gum disease were asked the number of times per week they used to floss before their diagnosis. The (incomplete) results are shown in the following table.

# Flossing per Week	Frequency	Relative Frequency	Cumulative Relative Freq.
0	27	0.4500	
1	18		
3			0.9333
6	3	0.0500	
7	1	0.0167	

Flossing Frequency for Adults with Gum Disease

a. Fill in the blanks in [link].
b. What percent of adults flossed six times per week?
c. What percent flossed at most three times per week?

Solution:

a.

# Flossing per Week	Frequency	Relative Frequency	Cumulative Relative Frequency
0	27	0.4500	0.4500
1	18	0.3000	0.7500
3	11	0.1833	0.9333
6	3	0.0500	0.9833
7	1	0.0167	1

b. 5.00%
c. 93.33%

Exercise:

Problem:

Nineteen immigrants to the U.S were asked how many years, to the nearest year, they have lived in the U.S. The data are as follows: 2 5 7 2 2 10 20 15 0 7 0 20 5 12 15 12 4 5 10 .

The following table was produced.

Data	Frequency	Relative Frequency	Cumulative Relative Frequency
0	2	$\frac{2}{19}$	0.1053
2	3	$\frac{3}{19}$	0.2632
4	1	$\frac{1}{19}$	0.3158
5	3	$\frac{3}{19}$	0.4737

Data	Frequency	Relative Frequency	Cumulative Relative Frequency
7	2	$\frac{2}{19}$	0.5789
10	2	$\frac{2}{19}$	0.6842
12	2	$\frac{2}{19}$	0.7895
15	1	$\frac{1}{19}$	0.8421
20	1	$\frac{1}{19}$	1.0000

Frequency of Immigrant Survey Responses

a. Fix the errors in [link]. Also, explain how someone might have arrived at the incorrect number(s).
b. Explain what is wrong with this statement: “47 percent of the people surveyed have lived in the U.S. for 5 years.”
c. Fix the statement in **b** to make it correct.
d. What fraction of the people surveyed have lived in the U.S. five or seven years?
e. What fraction of the people surveyed have lived in the U.S. at most 12 years?
f. What fraction of the people surveyed have lived in the U.S. fewer than 12 years?
g. What fraction of the people surveyed have lived in the U.S. from five to 20 years, inclusive?

Exercise:

Problem:

How much time does it take to travel to work? The following table shows the mean commute time by state for workers at least 16 years old who are not working at home. Find the mean travel time, and round off the answer properly.

24.0	24.3	25.9	18.9	27.5	17.9	21.8	20.9	16.7	27.3
18.2	24.7	20.0	22.6	23.9	18.0	31.4	22.3	24.0	25.5
24.7	24.6	28.1	24.9	22.6	23.6	23.4	25.7	24.8	25.5
21.2	25.7	23.1	23.0	23.9	26.0	16.3	23.1	21.4	21.5
27.0	27.0	18.6	31.7	23.3	30.1	22.9	23.3	21.7	18.6

Solution:

The sum of the travel times is 1,173.1. Divide the sum by 50 to calculate the mean value: 23.462. Because each state's travel time was measured to the nearest tenth, round this calculation to the nearest hundredth: 23.46.

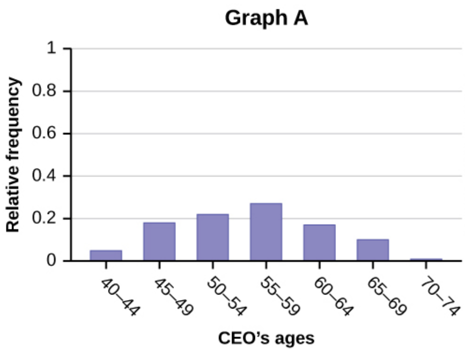
Exercise:

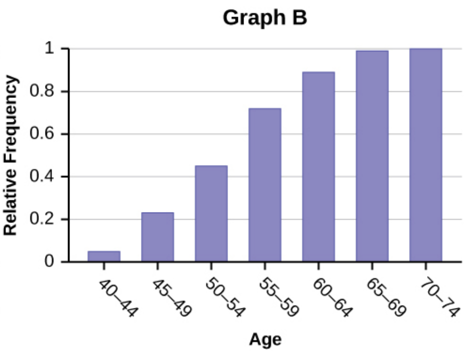
Problem:

Forbes magazine published data on the best small firms in 2012. These were firms which had been publicly traded for at least a year, have a stock price of at least $5 per share, and have reported annual revenue between $5 million and $1 billion. The following table shows the ages of the chief executive officers for the first 60 ranked firms.

Age	Frequency	Relative Frequency	Cumulative Relative Frequency
40–44	3		
45–49	11		
50–54	13		
55–59	16		
60–64	10		
65–69	6		
70–74	1		

a. What is the frequency for CEO ages between 54 and 65?
b. What percentage of CEOs are 65 years or older?
c. What is the relative frequency of ages under 50?
d. What is the cumulative relative frequency for CEOs younger than 55?
e. Which graph shows the relative frequency and which shows the cumulative relative frequency?





Use the following information to answer the next two exercises: The following table contains data on hurricanes that have made direct hits on the U.S. Between 1851 and 2004. A hurricane is given a strength category rating based on the minimum wind speed generated by the storm.

Category	Number of Direct Hits	Relative Frequency	Cumulative Frequency
1	109	0.3993	0.3993
2	72	0.2637	0.6630
3	71	0.2601	
4	18		0.9890
5	3	0.0110	1.0000
	Total = 273		

Frequency of Hurricane Direct Hits

Exercise:

Problem: What is the relative frequency of direct hits that were category 4 hurricanes?

a. 0.0768
b. 0.0659
c. 0.2601
d. Not enough information to calculate

Solution:

b

Exercise:

Problem:

What is the relative frequency of direct hits that were AT MOST a category 3 storm?

a. 0.3480
b. 0.9231
c. 0.2601
d. 0.3370

Glossary

Cumulative Relative Frequency

The term applies to an ordered set of observations from smallest to largest. The cumulative relative frequency is the sum of the relative frequencies for all values that are less than or equal to the given value.

Frequency

the number of times a value of the data occurs

Relative Frequency

the ratio of the number of times a value of the data occurs in the set of all outcomes to the number of all outcomes to the total number of outcomes

Experimental Design and Ethics

Does aspirin reduce the risk of heart attacks? Is one brand of fertilizer more effective at growing roses than another? Is fatigue as dangerous to a driver as the influence of alcohol? Questions like these are answered using randomized experiments. In this module, you will learn important aspects of experimental design. Proper study design ensures the production of reliable, accurate data.

The purpose of an experiment is to investigate the relationship between two variables. When one variable causes change in another, we call the first variable the **explanatory variable**. The affected variable is called the **response variable**. In a randomized experiment, the researcher manipulates values of the explanatory variable and measures the resulting changes in the response variable. The different values of the explanatory variable are called **treatments**. An **experimental unit** is a single object or individual to be measured.

You want to investigate the effectiveness of vitamin E in preventing disease. You recruit a group of subjects and ask them if they regularly take vitamin E. You notice that the subjects who take vitamin E exhibit better health on average than those who do not. Does this prove that vitamin E is effective in disease prevention? It does not. There are many differences between the two groups compared in addition to vitamin E consumption. People who take vitamin E regularly often take other steps to improve their health: exercise, diet, other vitamin supplements, choosing not to smoke. Any one of these factors could be influencing health. As described, this study does not prove that vitamin E is the key to disease prevention.

Additional variables that can cloud a study are called **lurking variables**. In order to prove that the explanatory variable is causing a change in the response variable, it is necessary to isolate the explanatory variable. The researcher must design her experiment in such a way that there is only one difference between groups being compared: the planned treatments. This is accomplished by the **random assignment** of experimental units to treatment groups. When subjects are assigned treatments randomly, all of the potential lurking variables are spread equally among the groups. At this point the only difference between groups is the one imposed by the researcher. Different outcomes measured in the response variable, therefore, must be a direct result of the different treatments. In this way, an experiment can prove a cause-and-effect connection between the explanatory and response variables.

The power of suggestion can have an important influence on the outcome of an experiment. Studies have shown that the expectation of the study participant can be as important as the actual medication. In one study of performance-enhancing drugs, researchers noted:

Results showed that believing one had taken the substance resulted in [performance*] times almost as fast as those associated with consuming the drug itself. In contrast, taking the drug without knowledge yielded no significant performance increment.*[footnote]

McClung, M. Collins, D. "Because I know it will!": placebo effects of an ergogenic aid on athletic performance. Journal of Sport & Exercise Psychology. 2007 Jun. 29(3):382-94. Web. April 30, 2013.

When participation in a study prompts a physical response from a participant, it is difficult to isolate the effects of the explanatory variable. To counter the power of suggestion, researchers set aside one treatment group as a **control group**. This group is given a **placebo** treatment–a treatment that cannot influence the response variable. The control group helps researchers balance the effects of being in an experiment with the effects of the active treatments. Of course, if you are participating in a study and you know that you are receiving a pill which contains no actual medication, then the power of suggestion is no longer a factor. **Blinding** in a randomized experiment preserves the power of suggestion. When a person involved in a research study is blinded, he does not know who is receiving the active treatment(s) and who is receiving the placebo treatment. A **double-blind**

experiment is one in which both the subjects and the researchers involved with the subjects are blinded.

Example:
Exercise:

Problem:

Researchers want to investigate whether taking aspirin regularly reduces the risk of heart attack. Four hundred men between the ages of 50 and 84 are recruited as participants. The men are divided randomly into two groups: one group will take aspirin, and the other group will take a placebo. Each man takes one pill each day for three years, but he does not know whether he is taking aspirin or the placebo. At the end of the study, researchers count the number of men in each group who have had heart attacks.

Identify the following values for this study: population, sample, experimental units, explanatory variable, response variable, treatments.

Solution:

The *population* is men aged 50 to 84.
The *sample* is the 400 men who participated.
The *experimental units* are the individual men in the study.
The *explanatory variable* is oral medication.
The *treatments* are aspirin and a placebo.
The *response variable* is whether a subject had a heart attack.

Example:
Exercise:

Problem:

The Smell & Taste Treatment and Research Foundation conducted a study to investigate whether smell can affect learning. Subjects completed mazes multiple times while wearing masks. They completed the pencil and paper mazes three times wearing floral-scented masks, and three times with unscented masks. Participants were assigned at random to wear the floral mask during the first three trials or during the last three trials. For each trial, researchers recorded the time it took to complete the maze and the subject's impression of the mask's scent: positive, negative, or neutral.

a. Describe the explanatory and response variables in this study.
b. What are the treatments?
c. Identify any lurking variables that could interfere with this study.
d. Is it possible to use blinding in this study?

Solution:

a. The explanatory variable is scent, and the response variable is the time it takes to complete the maze.
b. There are two treatments: a floral-scented mask and an unscented mask.
c. All subjects experienced both treatments. The order of treatments was randomly assigned so there were no differences between the treatment groups. Random assignment eliminates the problem of lurking variables.
d. Subjects will clearly know whether they can smell flowers or not, so subjects cannot be blinded in this study. Researchers timing the mazes can be blinded, though. The researcher who is observing a subject will not know which mask is being worn.

Example:

Exercise:

Problem:

A researcher wants to study the effects of birth order on personality. Explain why this study could not be conducted as a randomized experiment. What is the main problem in a study that cannot be designed as a randomized experiment?

Solution:

The explanatory variable is birth order. You cannot randomly assign a person's birth order. Random assignment eliminates the impact of lurking variables. When you cannot assign subjects to treatment groups at random, there will be differences between the groups other than the explanatory variable.

Note:
Try It
Exercise:

Problem:

You are concerned about the effects of texting on driving performance. Design a study to test the response time of drivers while texting and while driving only. How many seconds does it take for a driver to respond when a leading car hits the brakes?

a. Describe the explanatory and response variables in the study.
b. What are the treatments?
c. What should you consider when selecting participants?
d. Your research partner wants to divide participants randomly into two groups: one to drive without distraction and one to text and drive simultaneously. Is this a good idea? Why or why not?
e. Identify any lurking variables that could interfere with this study.
f. How can blinding be used in this study?

Solution:

a. Explanatory: presence of distraction from texting; response: response time measured in seconds
b. Driving without distraction and driving while texting
c. Answers will vary. Possible responses: Do participants regularly send and receive text messages? How long has the subject been driving? What is the age of the participants? Do participants have similar texting and driving experience?
d. This is not a good plan because it compares drivers with different abilities. It would be better to assign both treatments to each participant in random order.
e. Possible responses include: texting ability, driving experience, type of phone.
f. The researchers observing the trials and recording response time could be blinded to the treatment being applied.

Ethics

The widespread misuse and misrepresentation of statistical information often gives the field a bad name. Some say that “numbers don’t lie,” but the people who use numbers to support their claims often do.

A recent investigation of famous social psychologist, Diederik Stapel, has led to the retraction of his articles from some of the world’s top journals including *Journal of Experimental Social Psychology, Social Psychology, Basic and Applied Social Psychology*, *British Journal of Social Psychology*, and the magazine *Science*. Diederik Stapel is a former professor at Tilburg University in the Netherlands. Over the past two years, an extensive investigation involving three universities where Stapel has worked concluded that the psychologist is guilty of fraud on a colossal scale. Falsified data taints over 55 papers he authored and 10 Ph.D. dissertations that he supervised.

> *Stapel did not deny that his deceit was driven by ambition. But it was more complicated than that, he told me. He insisted that he loved social psychology but had been frustrated by the messiness of experimental data, which rarely led to clear conclusions. His lifelong obsession with elegance and order, he said, led him to concoct sexy results that journals found attractive. "It was a quest for aesthetics, for beauty—instead of the truth," he said. He described his behavior as an addiction that drove him to carry out acts of increasingly daring fraud, like a junkie seeking a bigger and better high.[footnote]*

Yudhijit Bhattacharjee, "The Mind of a Con Man," Magazine, New York Times, April 26, 2013. Available online at: http://www.nytimes.com/2013/04/28/magazine/diederik-stapels-audacious-academic-fraud.html?src=dayp&_r=2& (accessed May 1, 2013).

The committee investigating Stapel concluded that he is guilty of several practices including:

- creating datasets, which largely confirmed the prior expectations,
- altering data in existing datasets,
- changing measuring instruments without reporting the change, and
- misrepresenting the number of experimental subjects.

Clearly, it is never acceptable to falsify data the way this researcher did. Sometimes, however, violations of ethics are not as easy to spot.

Researchers have a responsibility to verify that proper methods are being followed. The report describing the investigation of Stapel's fraud states that, "statistical flaws frequently revealed a lack of familiarity with elementary statistics."[footnote] Many of Stapel's co-authors should have spotted irregularities in his data. Unfortunately, they did not know very much about statistical analysis, and they simply trusted that he was collecting and reporting data properly.

“Flawed Science: The Fraudulent Research Practices of Social Psychologist Diederik Stapel,” Tillburg University, November 28, 2012, http://www.tilburguniversity.edu/upload/064a10cd-bce5-4385-b9ff-05b840caeae6_120695_Rapp_nov_2012_UK_web.pdf (accessed May 1, 2013).

Many types of statistical fraud are difficult to spot. Some researchers simply stop collecting data once they have just enough to prove what they had hoped to prove. They don’t want to take the chance that a more extensive study would complicate their lives by producing data contradicting their hypothesis.

Professional organizations, like the American Statistical Association, clearly define expectations for researchers. There are even laws in the federal code about the use of research data.

When a statistical study uses human participants, as in medical studies, both ethics and the law dictate that researchers should be mindful of the safety of their research subjects. The U.S. Department of Health and Human Services oversees federal regulations of research studies with the aim of protecting participants. When a university or other research institution engages in research, it must ensure the safety of all human subjects. For this reason, research institutions establish oversight committees known as **Institutional Review Boards (IRB)**. All planned studies must be approved in advance by the IRB. Key protections that are mandated by law include the following:

- Risks to participants must be minimized and reasonable with respect to projected benefits.
- Participants must give **informed consent**. This means that the risks of participation must be clearly explained to the subjects of the study. Subjects must consent in writing, and researchers are required to keep documentation of their consent.
- Data collected from individuals must be guarded carefully to protect their privacy.

These ideas may seem fundamental, but they can be very difficult to verify in practice. Is removing a participant’s name from the data record sufficient to protect privacy? Perhaps the person’s identity could be discovered from the data that remains. What happens if the study does not proceed as planned and risks arise that were not anticipated? When is informed consent really necessary? Suppose your doctor wants a blood sample to check your cholesterol level. Once the sample has been tested, you expect the lab to dispose of the remaining blood. At that point the blood becomes biological waste. Does a researcher have the right to take it for use in a study?

It is important that students of statistics take time to consider the ethical questions that arise in statistical studies. How prevalent is fraud in statistical studies? You might be surprised—and disappointed. There is a website (www.retractionwatch.com) dedicated to cataloging retractions of study articles that have been proven fraudulent. A quick glance will show that the misuse of statistics is a bigger problem than most people realize.

Vigilance against fraud requires knowledge. Learning the basic theory of statistics will empower you to analyze statistical studies critically.

Example:
Exercise:

Problem:

Describe the unethical behavior in each example and describe how it could impact the reliability of the resulting data. Explain how the problem should be corrected.

A researcher is collecting data in a community.

a. She selects a block where she is comfortable walking because she knows many of the people living on the street.
b. No one seems to be home at four houses on her route. She does not record the addresses and does not return at a later time to try

to find residents at home.

c. She skips four houses on her route because she is running late for an appointment. When she gets home, she fills in the forms by selecting random answers from other residents in the neighborhood.

Solution:

a. By selecting a convenient sample, the researcher is intentionally selecting a sample that could be biased. Claiming that this sample represents the community is misleading. The researcher needs to select areas in the community at random.
b. Intentionally omitting relevant data will create bias in the sample. Suppose the researcher is gathering information about jobs and child care. By ignoring people who are not home, she may be missing data from working families that are relevant to her study. She needs to make every effort to interview all members of the target sample.
c. It is never acceptable to fake data. Even though the responses she uses are “real” responses provided by other participants, the duplication is fraudulent and can create bias in the data. She needs to work diligently to interview everyone on her route.

Note:
Try It
Exercise:

Problem:

Describe the unethical behavior, if any, in each example and describe how it could impact the reliability of the resulting data. Explain how the problem should be corrected.

A study is commissioned to determine the favorite brand of fruit juice among teens in California.

a. The survey is commissioned by the seller of a popular brand of apple juice.
b. There are only two types of juice included in the study: apple juice and cranberry juice.
c. Researchers allow participants to see the brand of juice as samples are poured for a taste test.
d. Twenty-five percent of participants prefer Brand X, 33% prefer Brand Y and 42% have no preference between the two brands. Brand X references the study in a commercial saying “Most teens like Brand X as much as or more than Brand Y.”

Solution:

a. This is not necessarily a problem. The study should be monitored carefully, however, to ensure that the company is not pressuring researchers to return biased results.
b. If the researchers truly want to determine the favorite brand of juice, then researchers should ask teens to compare different brands of the same type of juice. Choosing a sweet juice to compare against a sharp-flavored juice will not lead to an accurate comparison of brand quality.
c. Participants could be biased by the knowledge. The results may be different from those obtained in a blind taste test.
d. The commercial tells the truth, but not the whole truth. It leads consumers to believe that Brand X was preferred by more participants than Brand Y while the opposite is true.

References

“Vitamin E and Health,” Nutrition Source, Harvard School of Public Health, http://www.hsph.harvard.edu/nutritionsource/vitamin-e/ (accessed May 1, 2013).

Stan Reents. “Don’t Underestimate the Power of Suggestion,” athleteinme.com, http://www.athleteinme.com/ArticleView.aspx?id=1053 (accessed May 1, 2013).

Ankita Mehta. “Daily Dose of Aspiring Helps Reduce Heart Attacks: Study,” International Business Times, July 21, 2011. Also available online at http://www.ibtimes.com/daily-dose-aspirin-helps-reduce-heart-attacks-study-300443 (accessed May 1, 2013).

The Data and Story Library, http://lib.stat.cmu.edu/DASL/Stories/ScentsandLearning.html (accessed May 1, 2013).

M.L. Jacskon et al., “Cognitive Components of Simulated Driving Performance: Sleep Loss effect and Predictors,” Accident Analysis and Prevention Journal, Jan no. 50 (2013), http://www.ncbi.nlm.nih.gov/pubmed/22721550 (accessed May 1, 2013).

“Earthquake Information by Year,” U.S. Geological Survey. http://earthquake.usgs.gov/earthquakes/eqarchives/year/ (accessed May 1, 2013).

“Fatality Analysis Report Systems (FARS) Encyclopedia,” National Highway Traffic and Safety Administration. http://www-fars.nhtsa.dot.gov/Main/index.aspx (accessed May 1, 2013).

Data from www.businessweek.com (accessed May 1, 2013).

Data from www.forbes.com (accessed May 1, 2013).

“America’s Best Small Companies,” http://www.forbes.com/best-small-companies/list/ (accessed May 1, 2013).

U.S. Department of Health and Human Services, Code of Federal Regulations Title 45 Public Welfare Department of Health and Human Services Part 46 Protection of Human Subjects revised January 15, 2009. Section 46.111:Criteria for IRB Approval of Research.

“April 2013 Air Travel Consumer Report,” U.S. Department of Transportation, April 11 (2013), http://www.dot.gov/airconsumer/april-2013-air-travel-consumer-report (accessed May 1, 2013).

Lori Alden, “Statistics can be Misleading,” econoclass.com, http://www.econoclass.com/misleadingstats.html (accessed May 1, 2013).

Maria de los A. Medina, “Ethics in Statistics,” Based on “Building an Ethics Module for Business, Science, and Engineering Students” by Jose A. Cruz-Cruz and William Frey, Connexions, http://cnx.org/content/m15555/latest/ (accessed May 1, 2013).

Section Review

A poorly designed study will not produce reliable data. There are certain key components that must be included in every experiment. To eliminate lurking variables, subjects must be assigned randomly to different treatment groups. One of the groups must act as a control group, demonstrating what happens when the active treatment is not applied. Participants in the control group receive a placebo treatment that looks exactly like the active treatments but cannot influence the response variable. To preserve the integrity of the placebo, both researchers and subjects may be blinded. When a study is designed properly, the only difference between treatment groups is the one imposed by the researcher. Therefore, when groups respond differently to different treatments, the difference must be due to the influence of the explanatory variable.

“An ethics problem arises when you are considering an action that benefits you or some cause you support, hurts or reduces benefits to others, and violates some rule.”[footnote] Ethical violations in statistics are not always easy to spot. Professional associations and federal agencies post guidelines for proper conduct. It is important that you learn basic statistical procedures so that you can recognize proper data analysis.

Andrew Gelman, “Open Data and Open Methods,” Ethics and Statistics, http://www.stat.columbia.edu/~gelman/research/published/ChanceEthics1.pdf (accessed May 1, 2013).

Exercise:

Problem:

Design an experiment. Identify the explanatory and response variables. Describe the population being studied and the experimental units. Explain the treatments that will be used and how they will be assigned to the experimental units. Describe how blinding and placebos may be used to counter the power of suggestion.

Exercise:

Problem:

Discuss potential violations of the rule requiring informed consent.

a. Inmates in a correctional facility are offered good behavior credit in return for participation in a study.
b. A research study is designed to investigate a new children’s allergy medication.
c. Participants in a study are told that the new medication being tested is highly promising, but they are not told that only a small portion of participants will receive the new medication. Others will receive placebo treatments and traditional treatments.

Solution:

a. Inmates may not feel comfortable refusing participation, or may feel obligated to take advantage of the promised benefits. They may not feel truly free to refuse participation.
b. Parents can provide consent on behalf of their children, but children are not competent to provide consent for themselves.

c. All risks and benefits must be clearly outlined. Study participants must be informed of relevant aspects of the study in order to give appropriate consent.

HOMEWORK

Exercise:

Problem:

How does sleep deprivation affect your ability to drive? A recent study measured the effects on 19 professional drivers. Each driver participated in two experimental sessions: one after normal sleep and one after 27 hours of total sleep deprivation. The treatments were assigned in random order. In each session, performance was measured on a variety of tasks including a driving simulation.

Use key terms from this module to describe the design of this experiment.

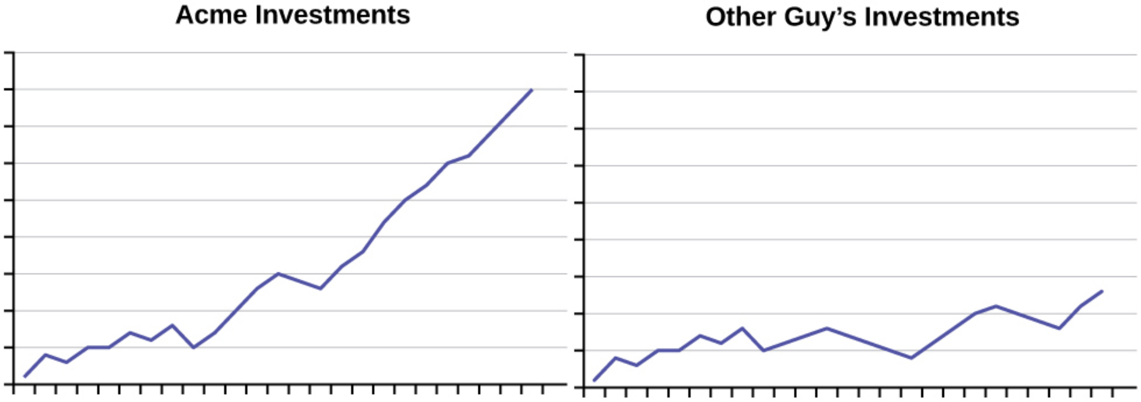
Solution:

Explanatory variable: amount of sleep
Response variable: performance measured in assigned tasks
Treatments: normal sleep and 27 hours of total sleep deprivation
Experimental Units: 19 professional drivers
Lurking variables: none – all drivers participated in both treatments
Random assignment: treatments were assigned in random order; this eliminated the effect of any “learning” that may take place during the first experimental session
Control/Placebo: completing the experimental session under normal sleep conditions
Blinding: researchers evaluating subjects’ performance must not know which treatment is being applied at the time

Exercise:

Problem:

An advertisement for Acme Investments displays the two graphs in the figure below to show the value of Acme’s product in comparison with the Other Guy’s product. Describe the potentially misleading visual effect of these comparison graphs. How can this be corrected?


Acme Investments

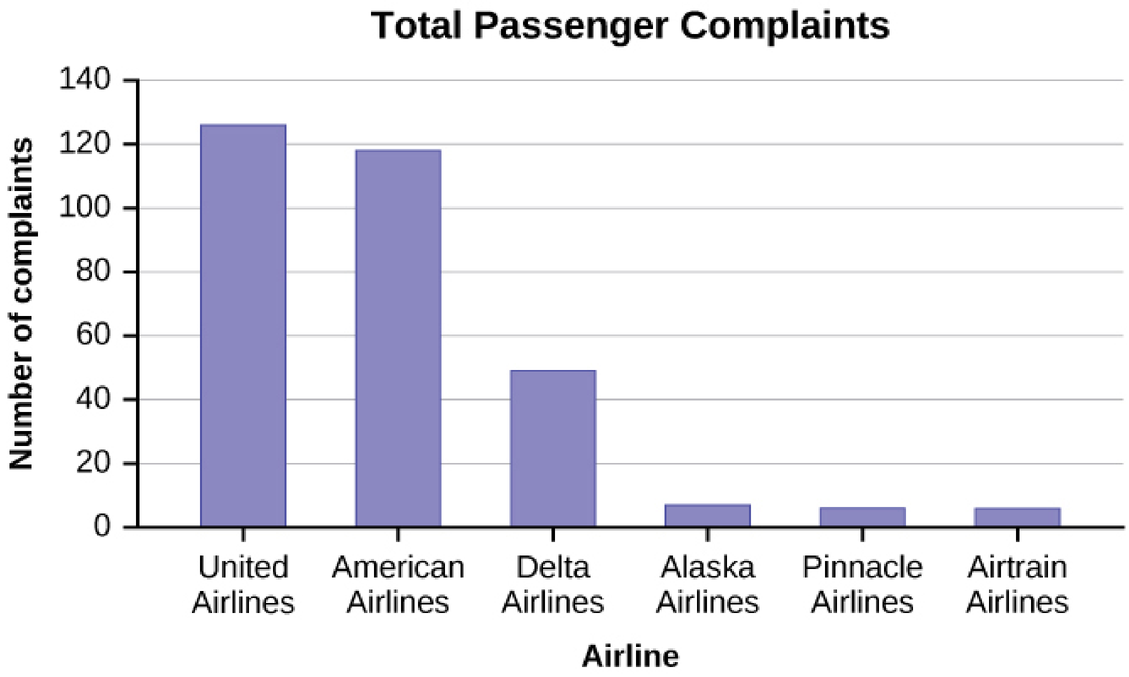
Other Guy’s Investments

As the graphs show, Acme consistently outperforms the Other Guys!

Exercise:

Problem:

The graph in the figure below shows the number of complaints for six different airlines as reported to the US Department of Transportation in February 2013. Alaska, Pinnacle, and Airtran Airlines have far fewer complaints reported than American, Delta, and United. Can we conclude that American, Delta, and United are the worst airline carriers since they have the most complaints?



Solution:

You cannot assume that the numbers of complaints reflect the quality of the airlines. The airlines shown with the greatest number of complaints are the ones with the most passengers. You must consider the appropriateness of methods for presenting data; in this case displaying totals is misleading.

Glossary

Explanatory Variable
: the independent variable in an experiment; the value controlled by researchers

Treatments
: different values or components of the explanatory variable applied in an experiment

Response Variable

the dependent variable in an experiment; the value that is measured for change at the end of an experiment

Experimental Unit
any individual or object to be measured

Lurking Variable
a variable that has an effect on a study even though it is neither an explanatory variable nor a response variable

Random Assignment
the act of organizing experimental units into treatment groups using random methods

Control Group
a group in a randomized experiment that receives an inactive treatment but is otherwise managed exactly as the other groups

Informed Consent
Any human subject in a research study must be cognizant of any risks or costs associated with the study. The subject has the right to know the nature of the treatments included in the study, their potential risks, and their potential benefits. Consent must be given freely by an informed, fit participant.

Placebo
an inactive treatment that has no real effect on the explanatory variable

Blinding
not telling participants which treatment a subject is receiving

Double-blinding
the act of blinding both the subjects of an experiment and the researchers who work with the subjects

Descriptive Statistics: Introduction class="introduction"

When you have large amounts of data, you will need to organize it in a way that makes sense. These ballots from an election are rolled together with similar ballots to keep them organized. (credit: William Greeson)

Note:

Chapter Objectives

By the end of this chapter, the student should be able to:

- Display data graphically and interpret graphs: stemplots, histograms, and box plots.
- Recognize, describe, and calculate the measures of location of data: quartiles and percentiles.
- Recognize, describe, and calculate the measures of the center of data: mean, median, and mode.
- Recognize, describe, and calculate the measures of the spread of data: variance, standard deviation, and range.

Once you have collected data, what will you do with it? Data can be described and presented in many different formats. For example, suppose you are interested in buying a house in a particular area. You may have no clue about the house prices, so you might ask your real estate agent to give you a sample data set of prices. Looking at all the prices in the sample often is overwhelming. A better way might be to look at the median price and the variation of prices. The median and variation are just two ways that you will learn to describe data. Your agent might also provide you with a graph of the data.

In this chapter, you will study numerical and graphical ways to describe and display your data. This area of statistics is called **"Descriptive Statistics."** You will learn how to calculate, and even more importantly, how to interpret these measurements and graphs.

A statistical graph is a tool that helps you learn about the shape or distribution of a sample or a population. A graph can be a more effective way of presenting data than a mass of numbers because we can see where data clusters and where there are only a few data values. Newspapers and the Internet use graphs to show trends and to enable readers to compare facts and figures quickly. Statisticians often graph data first to get a picture of the data. Then, more formal tools may be applied.

Some of the types of graphs that are used to summarize and organize data are the dot plot, the bar graph, the histogram, the stem-and-leaf plot, the frequency polygon (a type of broken line graph), the pie chart, and the box plot. In this chapter, we will briefly look at stem-and-leaf plots, line graphs, and bar graphs, as well as frequency polygons, and time series graphs. Our emphasis will be on histograms and box plots.

Note:
NOTE
This book contains instructions for constructing a histogram and a box plot for the TI-83+ and TI-84 calculators. The Texas Instruments (TI) website provides additional instructions for using these calculators.

Stem-and-Leaf Graphs (Stemplots), Line Graphs, and Bar Graphs

One simple graph, the **stem-and-leaf graph** or **stemplot**, comes from the field of exploratory data analysis. It is a good choice when the data sets are small. To create the plot, divide each observation of data into a stem and a leaf. The leaf consists of a **final significant digit**. For example, 23 has stem two and leaf three. The number 432 has stem 43 and leaf two. Likewise, the number 5,432 has stem 543 and leaf two. The decimal 9.3 has stem nine and leaf three. Write the stems in a vertical line from smallest to largest. Draw a vertical line to the right of the stems. Then write the leaves in increasing order next to their corresponding stem.

Example:
For Susan Dean's spring pre-calculus class, scores for the first exam were as follows (smallest to largest):
33; 42; 49; 49; 53; 55; 55; 61; 63; 67; 68; 68; 69; 69; 72; 73; 74; 78; 80; 83; 88; 88; 88; 90; 92; 94; 94; 94; 94; 96; 100

Stem	**Leaf**
3	3
4	2 9 9
5	3 5 5
6	1 3 7 8 8 9 9
7	2 3 4 8
8	0 3 8 8 8
9	0 2 4 4 4 4 6
10	0

Stem-and-Leaf Graph

The stemplot shows that most scores fell in the 60s, 70s, 80s, and 90s. Eight out of the 31 scores or approximately 26% $\left(\frac{8}{31}\right)$ were in the 90s or 100, a fairly high number of As.

Note:

Try It

Exercise:

Problem:

For the Park City basketball team, scores for the last 30 games were as follows (smallest to largest):

32; 32; 33; 34; 38; 40; 42; 42; 43; 44; 46; 47; 47; 48; 48; 48; 49; 50; 50; 51; 52; 52; 52; 53; 54; 56; 57; 57; 60; 61

Construct a stem plot for the data.

Solution:

Stem	**Leaf**
3	2 2 3 4 8
4	0 2 2 3 4 6 7 7 8 8 8 9
5	0 0 1 2 2 2 3 4 6 7 7
6	0 1

The stemplot is a quick way to graph data and gives an exact picture of the data. You want to look for an overall pattern and any outliers. An **outlier** is an observation of data that does not fit the rest of the data. It is sometimes called an **extreme value.** When you graph an outlier, it will appear not to fit the pattern of the graph. Some outliers are due to mistakes (for example, writing down 50 instead of 500) while others may indicate that something unusual is happening. It takes some background information to explain outliers, so we will cover them in more detail later.

Example:

The data are the distances (in kilometers) from a home to local supermarkets. Create a stemplot using the data:

1.1; 1.5; 2.3; 2.5; 2.7; 3.2; 3.3; 3.3; 3.5; 3.8; 4.0; 4.2; 4.5; 4.5; 4.7; 4.8; 5.5; 5.6; 6.5; 6.7; 12.3

Exercise:

Problem: Do the data seem to have any concentration of values?

Note: The leaves are to the right of the decimal.

Solution:

The value 12.3 may be an outlier. Values appear to concentrate at three and four kilometers.

Stem	**Leaf**
1	1 5
2	3 5 7
3	2 3 3 5 8
4	0 2 5 5 7 8
5	5 6
6	5 7
7	
8	
9	
10	
11	
12	3

Note:
Try It
Exercise:

Problem:

The following data show the distances (in miles) from the homes of off-campus statistics students to the college. Create a stem plot using the data and identify any outliers:

0.5; 0.7; 1.1; 1.2; 1.2; 1.3; 1.3; 1.5; 1.5; 1.7; 1.7; 1.8; 1.9; 2.0; 2.2; 2.5; 2.6; 2.8; 2.8; 2.8; 3.5; 3.8; 4.4; 4.8; 4.9; 5.2; 5.5; 5.7; 5.8; 8.0

Solution:

Stem	**Leaf**
0	5 7
1	1 2 2 3 3 5 5 7 7 8 9
2	0 2 5 6 8 8 8
3	5 8
4	4 8 9
5	2 5 7 8
6	
7	
8	0

The value 8.0 may be an outlier. Values appear to concentrate at one and two miles.

Example:
Exercise:

Problem:

A **side-by-side stem-and-leaf plot** allows a comparison of the two data sets in two columns. In a side-by-side stem-and-leaf plot, two sets of leaves share the same stem. The leaves are to the left and the right of the stems. [link] and [link] show the ages of presidents at their inauguration and at their death. Construct a side-by-side stem-and-leaf plot using this data.

Solution:

Ages at Inauguration		Ages at Death
9 9 8 7 7 7 6 3 2	4	6 9
8 7 7 7 7 6 6 6 5 5 5 5 4 4 4 4 4 2 1 1 1 1 1 0	5	3 6 6 7 7 8
9 5 4 4 2 1 1 1 0	6	0 0 3 3 4 4 5 6 7 7 7 8
	7	0 0 1 1 1 4 7 8 8 9
	8	0 1 3 5 8
	9	0 0 3 3

President	Age	President	Age	President	Age
Washington	57	Lincoln	52	Hoover	54
J. Adams	61	A. Johnson	56	F. Roosevelt	51
Jefferson	57	Grant	46	Truman	60
Madison	57	Hayes	54	Eisenhower	62
Monroe	58	Garfield	49	Kennedy	43
J. Q. Adams	57	Arthur	51	L. Johnson	55

President	**Age**	**President**	**Age**	**President**	**Age**
Jackson	61	Cleveland	47	Nixon	56
Van Buren	54	B. Harrison	55	Ford	61
W. H. Harrison	68	Cleveland	55	Carter	52
Tyler	51	McKinley	54	Reagan	69
Polk	49	T. Roosevelt	42	G.H.W. Bush	64
Taylor	64	Taft	51	Clinton	47
Fillmore	50	Wilson	56	G. W. Bush	54
Pierce	48	Harding	55	Obama	47
Buchanan	65	Coolidge	51		

Presidential Ages at Inauguration

President	**Age**	**President**	**Age**	**President**	**Age**
Washington	67	Lincoln	56	Hoover	90
J. Adams	90	A. Johnson	66	F. Roosevelt	63
Jefferson	83	Grant	63	Truman	88
Madison	85	Hayes	70	Eisenhower	78
Monroe	73	Garfield	49	Kennedy	46
J. Q. Adams	80	Arthur	56	L. Johnson	64
Jackson	78	Cleveland	71	Nixon	81
Van Buren	79	B. Harrison	67	Ford	93
W. H. Harrison	68	Cleveland	71	Reagan	93

President	Age	President	Age	President	Age
Tyler	71	McKinley	58		
Polk	53	T. Roosevelt	60		
Taylor	65	Taft	72		
Fillmore	74	Wilson	67		
Pierce	64	Harding	57		
Buchanan	77	Coolidge	60		

Presidential Age at Death

Note:
Exercise:

Problem:

The table shows the number of wins and losses the Atlanta Hawks have had in 42 seasons. Create a side-by-side stem-and-leaf plot of these wins and losses.

Losses	Wins	Year	Losses	Wins	Year
34	48	1968–1969	41	41	1989–1990
34	48	1969–1970	39	43	1990–1991
46	36	1970–1971	44	38	1991–1992
46	36	1971–1972	39	43	1992–1993
36	46	1972–1973	25	57	1993–1994
47	35	1973–1974	40	42	1994–1995
51	31	1974–1975	36	46	1995–1996
53	29	1975–1976	26	56	1996–1997

Losses	Wins	Year	Losses	Wins	Year
51	31	1976–1977	32	50	1997–1998
41	41	1977–1978	19	31	1998–1999
36	46	1978–1979	54	28	1999–2000
32	50	1979–1980	57	25	2000–2001
51	31	1980–1981	49	33	2001–2002
40	42	1981–1982	47	35	2002–2003
39	43	1982–1983	54	28	2003–2004
42	40	1983–1984	69	13	2004–2005
48	34	1984–1985	56	26	2005–2006
32	50	1985–1986	52	30	2006–2007
25	57	1986–1987	45	37	2007–2008
32	50	1987–1988	35	47	2008–2009
30	52	1988–1989	29	53	2009–2010

Solution:

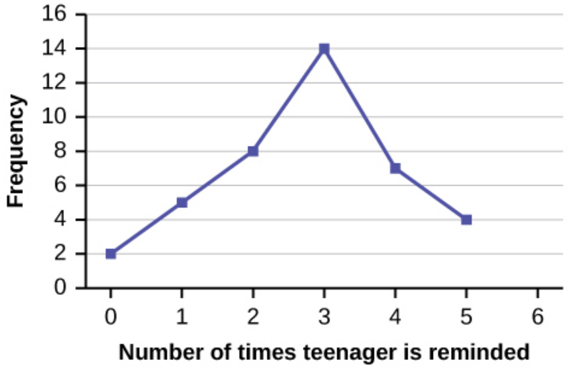
Atlanta Hawks Wins and Losses		
Number of Wins		**Number of Losses**
3	1	9
9 8 8 6 5	2	5 5 9
8 7 6 6 5 5 4 3 1 1 1 1 0	3	0 2 2 2 2 4 4 5 6 6 6 9 9 9
8 8 7 6 6 6 3 3 3 2 2 1 1 0	4	0 0 1 1 2 4 5 6 6 7 7 8 9

Atlanta Hawks Wins and Losses		
Number of Wins		**Number of Losses**
7 7 6 3 2 0 0 0 0	5	1 1 1 2 3 4 4 6 7
	6	9

Another type of graph that is useful for specific data values is a **line graph**. In the particular line graph shown in [link], the ***x*-axis** (horizontal axis) consists of **data values** and the ***y*-axis** (vertical axis) consists of **frequency points**. The frequency points are connected using line segments.

Example:
In a survey, 40 mothers were asked how many times per week a teenager must be reminded to do his or her chores. The results are shown in [link] and in [link].

Number of times teenager is reminded	**Frequency**
0	2
1	5
2	8
3	14
4	7
5	4



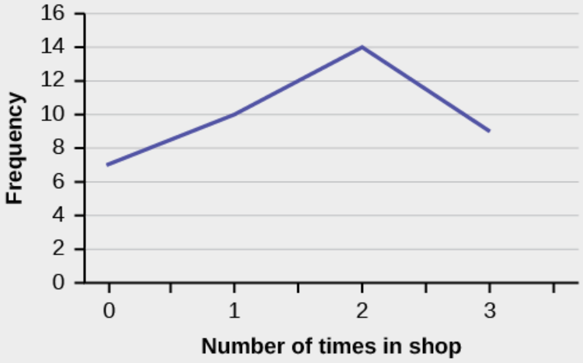
Note:
Try It
Exercise:

Problem:

In a survey, 40 people were asked how many times per year they had their car in the shop for repairs. The results are shown in the following table. Construct a line graph.

Number of times in shop	**Frequency**
0	7
1	10
2	14
3	9

Solution:



Bar graphs consist of bars that are separated from each other. The bars can be rectangles or they can be rectangular boxes (used in three-dimensional plots), and they can be vertical or horizontal. The **bar graph** shown in [link] has age groups represented on the ***x*-axis** and proportions on the ***y*-axis**.

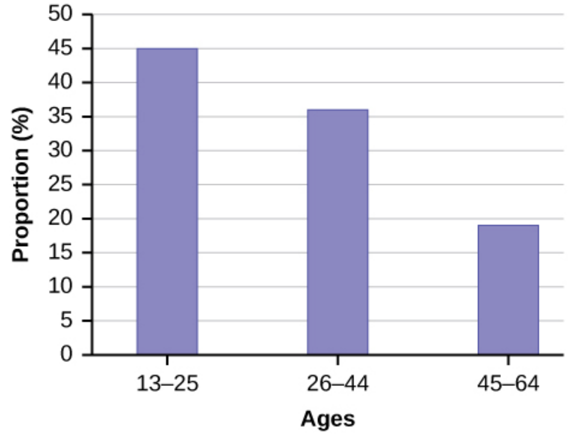
Example:
Exercise:

Problem:

By the end of 2011, Facebook had over 146 million users in the United States. The following table shows three age groups, the number of users in each age group, and the proportion (%) of users in each age group. Construct a bar graph using this data.

Age groups	Number of Facebook users	Proportion (%) of Facebook users
13–25	65,082,280	45%
26–44	53,300,200	36%
45–64	27,885,100	19%

Solution:



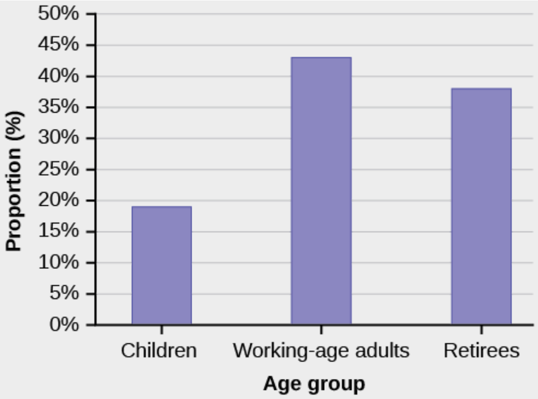
Note:
Try It
Exercise:

Problem:

The population in Park City is made up of children, working-age adults, and retirees. The following table shows the three age groups, the number of people in the town from each age group, and the proportion (%) of people in each age group. Construct a bar graph showing the proportions.

Age groups	Number of people	Proportion of population
Children	67,059	19%
Working-age adults	152,198	43%
Retirees	131,662	38%

Solution:



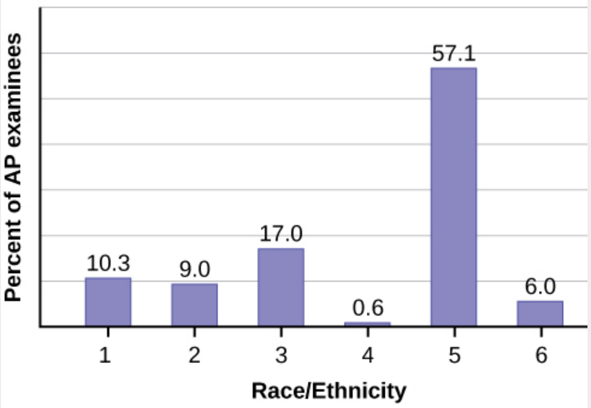
Example:
Exercise:

Problem:

The columns in the following axis contain: the race or ethnicity of students in U.S. Public Schools for the class of 2011, percentages for the Advanced Placement examine population for that class, and percentages for the overall student population. Create a bar graph with the student race or ethnicity (qualitative data) on the x-axis, and the Advanced Placement examinee population percentages on the y-axis.

Race/Ethnicity	AP Examinee Population	Overall Student Population
1 = Asian, Asian American or Pacific Islander	10.3%	5.7%
2 = Black or African American	9.0%	14.7%
3 = Hispanic or Latino	17.0%	17.6%
4 = American Indian or Alaska Native	0.6%	1.1%
5 = White	57.1%	59.2%
6 = Not reported/other	6.0%	1.7%

Solution:



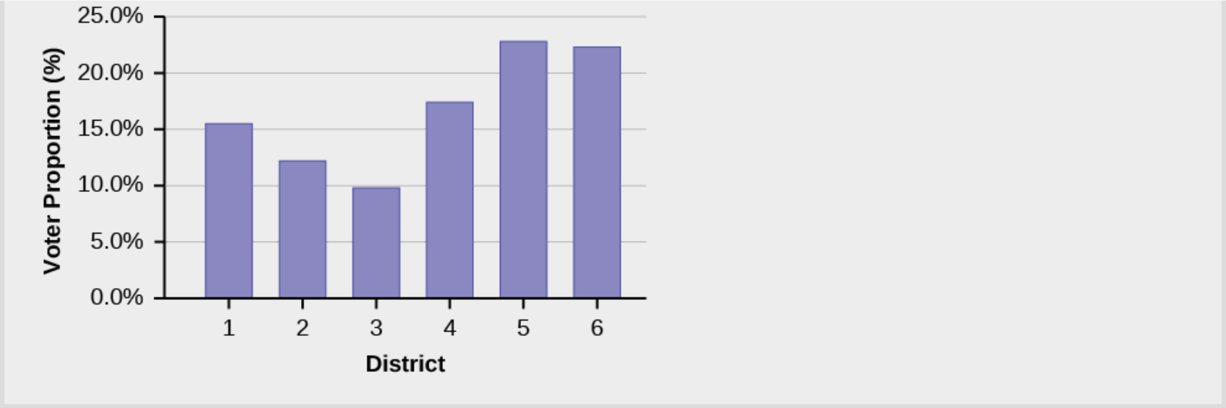
Note:
Try It
Exercise:

Problem:

Park city is broken down into six voting districts. The table shows the percent of the total registered voter population that lives in each district as well as the percent total of the entire population that lives in each district. Construct a bar graph that shows the registered voter population by district.

District	Registered voter population	Overall city population
1	15.5%	19.4%
2	12.2%	15.6%
3	9.8%	9.0%
4	17.4%	18.5%
5	22.8%	20.7%
6	22.3%	16.8%

Solution:



References

Burbary, Ken. *Facebook Demographics Revisited – 2001 Statistics*, 2011. Available online at http://www.kenburbary.com/2011/03/facebook-demographics-revisited-2011-statistics-2/ (accessed August 21, 2013).

“9th Annual AP Report to the Nation.” CollegeBoard, 2013. Available online at http://apreport.collegeboard.org/goals-and-findings/promoting-equity (accessed September 13, 2013).

“Overweight and Obesity: Adult Obesity Facts.” Centers for Disease Control and Prevention. Available online at http://www.cdc.gov/obesity/data/adult.html (accessed September 13, 2013).

Section Review

A **stem-and-leaf plot** is a way to plot data and look at the distribution. In a stem-and-leaf plot, all data values within a class are visible. The advantage in a stem-and-leaf plot is that all values are listed, unlike a histogram, which gives classes of data values. A **line graph** is often used to represent a set of data values in which a quantity varies with time. These graphs are useful for finding trends. That is, finding a general pattern in data sets including temperature, sales, employment, company profit or cost over a period of time. A **bar graph** is a chart that uses either horizontal or vertical bars to show comparisons among categories. One axis of the chart shows the specific categories being compared, and the other axis represents a discrete value. Some bar graphs present bars clustered in groups of more than one (grouped bar graphs), and others show the bars divided into subparts to show cumulative effect (stacked bar graphs). Bar graphs are especially useful when categorical data is being used.

For each of the following data sets, create a stem plot and identify any outliers.
Exercise:

Problem:

The miles per gallon rating for 30 cars are shown below (lowest to highest).
19, 19, 19, 20, 21, 21, 25, 25, 25, 26, 26, 28, 29, 31, 31, 32, 32, 33, 34, 35, 36, 37, 37, 38, 38, 38, 38, 41, 43, 43

Solution:

Stem	Leaf
1	9 9 9
2	0 1 1 5 5 5 6 6 8 9
3	1 1 2 2 3 4 5 6 7 7 8 8 8 8
4	1 3 3

Exercise:

Problem:

The height in feet of 25 trees is shown below (lowest to highest).
25, 27, 33, 34, 34, 34, 35, 37, 37, 38, 39, 39, 39, 40, 41, 45, 46, 47, 49, 50, 50, 53, 53, 54, 54

Exercise:

Problem:

The data are the prices of different laptops at an electronics store. Round each value to the nearest ten.
249, 249, 260, 265, 265, 280, 299, 299, 309, 319, 325, 326, 350, 350, 350, 365, 369, 389, 409, 459, 489, 559, 569, 570, 610

Solution:

Stem	Leaf
2	5 5 6 7 7 8

Stem	**Leaf**
3	0 0 1 2 3 3 5 5 5 7 7 9
4	1 6 9
5	6 7 7
6	1

Exercise:

Problem:

The data are daily high temperatures in a town for one month.
61, 61, 62, 64, 66, 67, 67, 67, 68, 69, 70, 70, 70, 71, 71, 72, 74, 74, 74, 75, 75, 75, 76, 76, 77, 78, 78, 79, 79, 95

For the next three exercises, use the data to construct a line graph.

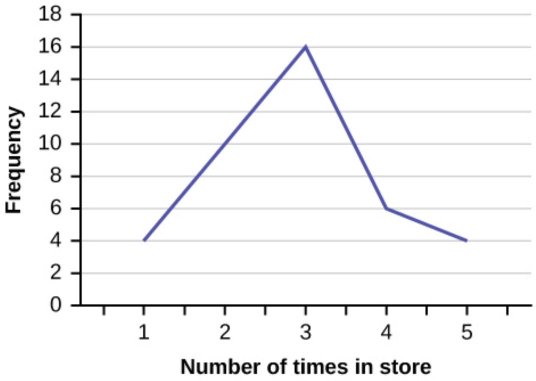
Exercise:

Problem:

In a survey, 40 people were asked how many times they visited a store before making a major purchase. The results are shown in the following table.

Number of times in store	**Frequency**
1	4
2	10
3	16
4	6
5	4

Solution:



Exercise:

Problem:

In a survey, several people were asked how many years it has been since they purchased a mattress. The results are shown in the following table.

Years since last purchase	**Frequency**
0	2
1	8
2	13
3	22
4	16
5	9

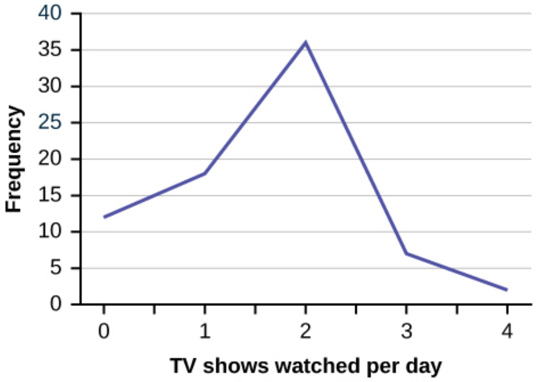
Exercise:

Problem:

Several children were asked how many TV shows they watch each day. The results of the survey are shown in the following table.

Number of TV Shows	**Frequency**
0	12
1	18
2	36
3	7
4	2

Solution:



Exercise:

Problem:

The students in Ms. Ramirez's math class have birthdays in each of the four seasons. The following table shows the four seasons, the number of students who have birthdays in each season, and the percentage (%) of students in each group. Construct a bar graph showing the number of students.

Seasons	**Number of students**	**Proportion of population**
Spring	8	24%
Summer	9	26%
Autumn	11	32%

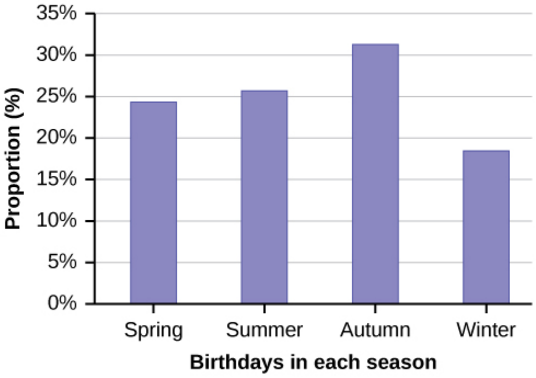
Seasons	Number of students	Proportion of population
Winter	6	18%

Exercise:

Problem:

Using the data from Mrs. Ramirez’s math class supplied in [link], construct a bar graph showing the percentages.

Solution:



Exercise:

Problem:

David County has six high schools. Each school sent students to participate in a county-wide science competition. The following table shows the percentage breakdown of competitors from each school, and the percentage of the entire student population of the county that goes to each school. Construct a bar graph that shows the population percentage of competitors from each school.

High School	Science competition population	Overall student population
Alabaster	28.9%	8.6%
Concordia	7.6%	23.2%
Genoa	12.1%	15.0%
Mocksville	18.5%	14.3%

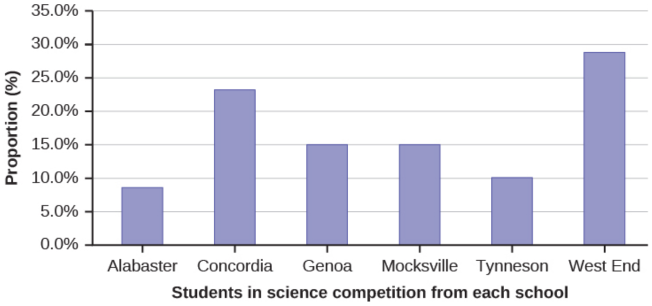
High School	Science competition population	Overall student population
Tynneson	24.2%	10.1%
West End	8.7%	28.8%

Exercise:

Problem:

Use the data from the David County science competition supplied in [link]. Construct a bar graph that shows the county-wide population percentage of students at each school.

Solution:



Homework

Exercise:

Problem: Student grades on a chemistry exam were: 77, 78, 76, 81, 86, 51, 79, 82, 84, 99

a. Construct a stem-and-leaf plot of the data.
b. Are there any potential outliers? If so, which scores are they? Why do you consider them outliers?

Exercise:

Problem:

The following table contains the 2010 obesity rates in U.S. states and Washington, DC.

State	Percent (%)	State	Percent (%)	State	Percent (%)
Alabama	32.2	Kentucky	31.3	North Dakota	27.2
Alaska	24.5	Louisiana	31.0	Ohio	29.2
Arizona	24.3	Maine	26.8	Oklahoma	30.4
Arkansas	30.1	Maryland	27.1	Oregon	26.8
California	24.0	Massachusetts	23.0	Pennsylvania	28.6
Colorado	21.0	Michigan	30.9	Rhode Island	25.5
Connecticut	22.5	Minnesota	24.8	South Carolina	31.5
Delaware	28.0	Mississippi	34.0	South Dakota	27.3
Washington, DC	22.2	Missouri	30.5	Tennessee	30.8
Florida	26.6	Montana	23.0	Texas	31.0
Georgia	29.6	Nebraska	26.9	Utah	22.5
Hawaii	22.7	Nevada	22.4	Vermont	23.2
Idaho	26.5	New Hampshire	25.0	Virginia	26.0
Illinois	28.2	New Jersey	23.8	Washington	25.5
Indiana	29.6	New Mexico	25.1	West Virginia	32.5
Iowa	28.4	New York	23.9	Wisconsin	26.3
Kansas	29.4	North Carolina	27.8	Wyoming	25.1

a. Use a random number generator to randomly pick eight states. Construct a bar graph of the obesity rates of those eight states.
b. Construct a bar graph for all the states beginning with the letter "A."

c. Construct a bar graph for all the states beginning with the letter "M."

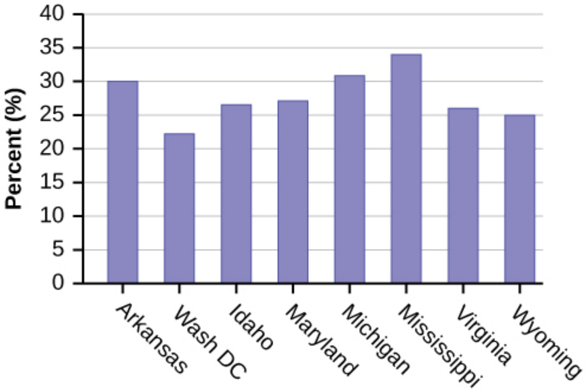
Solution:

a. Example solution for using the random number generator for the TI-84+ to generate a simple random sample of 8 states. Instructions are as follows.

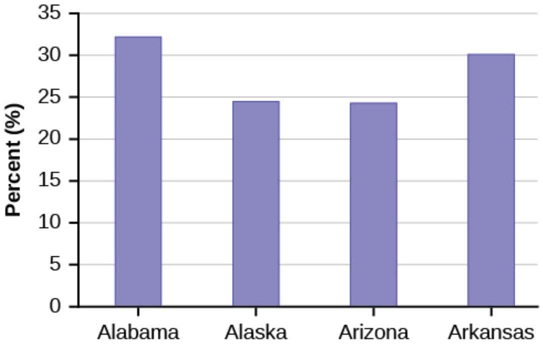
   - Number the entries in the table 1–51 (Includes Washington, DC; Numbered vertically)
   - Press MATH
   - Arrow over to PRB
   - Press 5:randInt(
   - Enter 51,1,8)

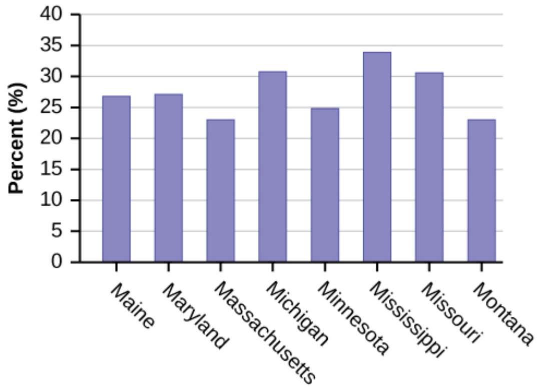
   Eight numbers are generated (use the right arrow key to scroll through the numbers). The numbers correspond to the numbered states (for this example: {47 21 9 23 51 13 25 4}. If any numbers are repeated, generate a different number by using 5:randInt(51,1)). Here, the states (and Washington DC) are {Arkansas, Washington DC, Idaho, Maryland, Michigan, Mississippi, Virginia, Wyoming}.

   Corresponding percents are {30.1, 22.2, 26.5, 27.1, 30.9, 34.0, 26.0, 25.1}.



b.





c.

Exercise:

Problem: Following are the 2010 obesity rates by U.S. states and Washington, DC.

State	Percent (%)	State	Percent (%)	State	Percent (%)
Alabama	32.2	Kentucky	31.3	North Dakota	27.2
Alaska	24.5	Louisiana	31.0	Ohio	29.2
Arizona	24.3	Maine	26.8	Oklahoma	30.4
Arkansas	30.1	Maryland	27.1	Oregon	26.8
California	24.0	Massachusetts	23.0	Pennsylvania	28.6
Colorado	21.0	Michigan	30.9	Rhode Island	25.5
Connecticut	22.5	Minnesota	24.8	South Carolina	31.5
Delaware	28.0	Mississippi	34.0	South Dakota	27.3
Washington, DC	22.2	Missouri	30.5	Tennessee	30.8
Florida	26.6	Montana	23.0	Texas	31.0

State	Percent (%)	State	Percent (%)	State	Percent (%)
Georgia	29.6	Nebraska	26.9	Utah	22.5
Hawaii	22.7	Nevada	22.4	Vermont	23.2
Idaho	26.5	New Hampshire	25.0	Virginia	26.0
Illinois	28.2	New Jersey	23.8	Washington	25.5
Indiana	29.6	New Mexico	25.1	West Virginia	32.5
Iowa	28.4	New York	23.9	Wisconsin	26.3
Kansas	29.4	North Carolina	27.8	Wyoming	25.1

Construct a bar graph of obesity rates of your state and the four states closest to your state. Hint: Label the x-axis with the states.

Histograms, Frequency Polygons, and Time Series Graphs

For most of the work you do in this book, you will use a histogram to display the data. One advantage of a histogram is that it can readily display large data sets. A rule of thumb is to use a histogram when the data set consists of 100 values or more.

A **histogram** consists of contiguous (adjoining) boxes. It has both a horizontal axis and a vertical axis. The horizontal axis is labeled with what the data represents (for instance, distance from your home to school). The vertical axis is labeled either **frequency** or **relative frequency** (or percent frequency or probability). The graph will have the same shape whether you choose to use frequency or relative frequency. The histogram (like the stemplot) can give you the shape of the data, the center, and the spread of the data.

The relative frequency is equal to the frequency for an observed value of the data divided by the total number of data values in the sample. (Remember, frequency is defined as the number of times an answer occurs.) If:

- f = frequency
- n = total number of data values (or the sum of the individual frequencies), and
- RF = relative frequency,

then:
Equation:

$$\mathrm{RF} = \frac{f}{n}$$

For example, if three students in Mr. Ahab's English class of 40 students received from 90% to 100%, then, f = 3, n = 40, and $RF = \frac{f}{n} = \frac{3}{40}$ = 0.075. 7.5% of the students received 90–100%. 90–100% are quantitative measures.

To construct a histogram, first decide how many **bars** or **class intervals**, also called classes, will best represent the data. Failing to use enough bars may not properly summarize the data, while using too many bars may be overly detailed. Many histograms consist of five to 15 bars or classes for clarity. Next, choose a starting point for the first interval to be less than or equal to the smallest data value and an ending value that is greater than the largest data value. Once the number of bars, starting value, and ending value have been decided, use these values to determine the width of each class interval and adjust as needed. The next two examples cover how to construct a histogram using continuous data and how to create a histogram using discrete data.

Example:
The following data are the heights (in inches to the nearest half inch) of 100 male semiprofessional soccer players. The heights are **continuous** data, since height is measured.

60; 60.5; 61; 61; 61.5
63.5; 63.5; 63.5
64; 64; 64; 64; 64; 64; 64; 64.5; 64.5; 64.5; 64.5; 64.5; 64.5; 64.5; 64.5
66; 66; 66; 66; 66; 66; 66; 66; 66; 66; 66.5; 66.5; 66.5; 66.5; 66.5; 66.5; 66.5; 66.5; 66.5; 66.5; 66.5; 67; 67; 67; 67; 67; 67; 67; 67; 67; 67; 67; 67; 67.5; 67.5; 67.5; 67.5; 67.5; 67.5; 67.5
68; 68; 69; 69; 69; 69; 69; 69; 69; 69; 69; 69; 69.5; 69.5; 69.5; 69.5; 69.5
70; 70; 70; 70; 70; 70; 70.5; 70.5; 70.5; 71; 71; 71
72; 72; 72; 72.5; 72.5; 73; 73.5
74
The smallest data value is 60 and since this is a nice number, we can use 60 as the starting value.
The largest value is 74. Thus, we can use 75 or 80, since this is also a nice number, as an ending value.

Next, calculate the width of each bar or class interval. To calculate this width, subtract the starting point from the ending value and divide by the number of bars (you must choose the number of bars you desire). Suppose you choose ten bars and decide to use 80 as the ending value.

Equation:

$$\frac{80-60}{10}=2$$

Note:

NOTE

A guideline that is followed by some for choosing the number of bars or class intervals to use is to take the square root of the number of data values and then round to the nearest whole number, if necessary. For example, if there are 150 values of data, take the square root of 150 and round to 12 bars or intervals.

The boundaries are:

- 60
- 60 + 2 = 62
- 62 + 2 = 64
- 64 + 2 = 66
- 66 + 2 = 68
- 68 + 2 = 70
- 70 + 2 = 72
- 72 + 2 = 74
- 74 + 2 = 76
- 76 + 2 = 78
- 78 + 2 = 80

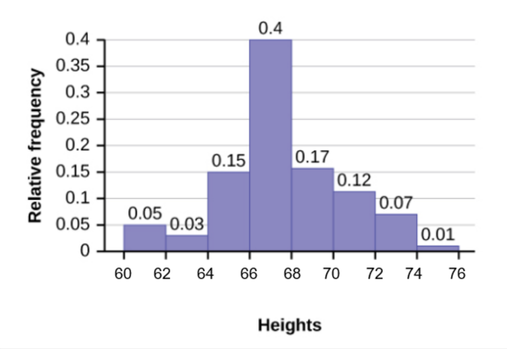
Note: Any data value that is equal to a boundary value should be placed in the bar to the **right** of that boundary. For example, the data value 74 will fall into the bar representing the interval from 74 to 76.

We then construct a relative frequency table and use that to draw the histogram.

CLASS INTERVAL	**FREQUENCY**	**RELATIVE FREQUENCY**
$60 \le x < 62$	5	$\frac{5}{100}$ or 0.05
$62 \le x < 64$	3	$\frac{3}{100}$ or 0.03
$64 \le x < 66$	15	$\frac{15}{100}$ or 0.15
$66 \le x < 68$	40	$\frac{40}{100}$ or 0.4

CLASS INTERVAL	FREQUENCY	RELATIVE FREQUENCY
$68 \le x < 70$	17	$\frac{17}{100}$ or 0.17
$70 \le x < 72$	12	$\frac{12}{100}$ or 0.12
$72 \le x < 74$	7	$\frac{7}{100}$ or 0.07
$74 \le x < 76$	1	$\frac{1}{100}$ or 0.01

The following histogram displays the heights on the x-axis and relative frequency on the y-axis.



Note:
Try It
Exercise:

Problem:

The following data are the shoe sizes of 50 male students. The sizes are continuous data since shoe size is measured. Construct a histogram and calculate the width of each bar or class interval. Suppose you choose six bars.

9; 9; 9.5; 9.5; 10; 10; 10; 10; 10; 10; 10.5; 10.5; 10.5; 10.5; 10.5; 10.5; 10.5; 10.5
11; 11; 11; 11; 11; 11; 11; 11; 11; 11; 11; 11; 11; 11.5; 11.5; 11.5; 11.5; 11.5; 11.5; 11.5; 12; 12; 12; 12; 12; 12; 12; 12.5; 12.5; 12.5; 12.5; 14

Solution:

Smallest value: 9

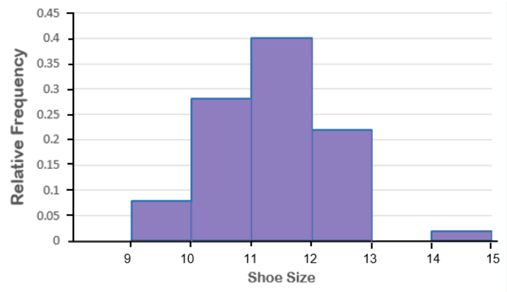
Largest value: 14

Choose 9 as the starting point and 15 as the ending value.

$\frac{15-9}{6} = 1$

The calculations suggests using 1 as the width of each bar or class interval.

Using class intervals with a width of one and relative frequency for the vertical axis, you should obtain the following histogram:

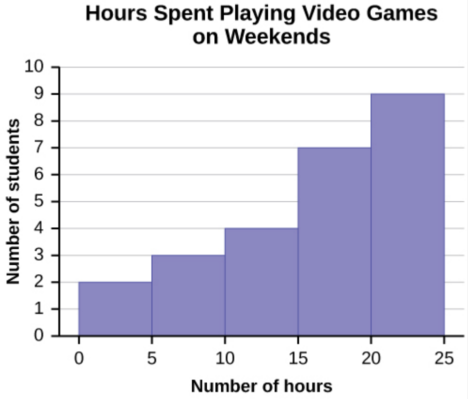


Example:
Exercise:

Problem: Using this data set, construct a histogram.

Number of Hours My Classmates Spent Playing Video Games on Weekends				
9.95	10	2.25	16.75	0
19.5	22.5	7.5	15	12.75
5.5	11	10	20.75	17.5
23	21.9	24	23.75	18
20	15	22.9	18.8	20.5

Solution:



Notice that for the histogram above, five class intervals are used, which comes from taking the square root of the number of data values (25). The smallest data value is 0 and the largest data value is 24, so the starting and ending values were chosen to be 0 and 25 respectively, which gives a width of five for the class intervals.

Note:
Creating a Histogram on the TI 83/84
Below are calculator instructions for entering data and creating a histogram.

- Press Y=. Press CLEAR to delete any equations.
- Press STAT 1:EDIT. If L1 has data in it, arrow up onto the name L1, press CLEAR and then arrow down. If necessary, do the same for L2.
- Into L1, enter the data values (once each).
- Into L2, enter the frequency corresponding to each data value.
- Press WINDOW. Set Xmin = determined starting value, Xscl = interval width, Xmax = determined ending value, Ymin = –1, Ymax = a number greater than the highest frequency, Yscl = 1, Xres = 1.
- Press 2nd Y=. Start by pressing 4:Plotsoff ENTER.
- Press 2nd Y=. Press 1:Plot1. Press ENTER. Arrow down to TYPE. Arrow to the 3rd picture (histogram). Press ENTER.
- Arrow down to Xlist: Enter L1 (2nd 1). Arrow down to Freq. Enter L2 (2nd 2).
- Press GRAPH.
- Use the TRACE key and the arrow keys to examine the histogram.

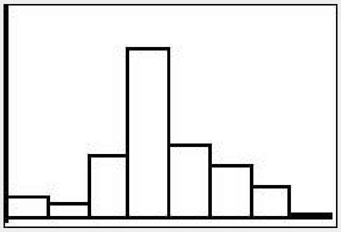
Note: You can also use Zoom Stat (#9 in the Zoom menu) and the calculator will automatically graph a histogram in a window that will give you a good summary of the data. You can then adjust the scales, if necessary.

Note:
Try It
Exercise:

Problem:

Construct a histogram of the data in [link] on your calculator. Notice how similar it looks to the one constructed by hand.

Solution:



The window for the above histogram was set to: Xmin = 60, Xmax = 76, Xscl = 2, Ymin = -1, and Ymax = 50.

Example:
The following data are the number of books bought by 50 part-time college students at ABC College. The number of books is **discrete data**, since books are counted.

1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1
2; 2; 2; 2; 2; 2; 2; 2; 2; 2

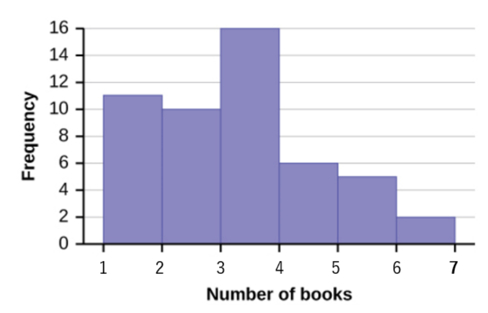
3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3
4; 4; 4; 4; 4; 4
5; 5; 5; 5; 5
6; 6
The data values range from 1 to 6, so we can use a starting value of 1 and an ending value of 7.

Next, determine the width of each bar or class interval. If the data are discrete, then the width should be a whole number. Since the data values in this case range from 1 to 6, it would make sense to use a width of one, which would give us 6 bars. We could also use a width of two or three, but this would give only 2 or 3 bars, which wouldn't give a very good summary of the data.

Organize the data into a frequency table as follows.

CLASS INTERVAL	**FREQUENCY**
$1 \le x < 2$	11
$2 \le x < 3$	10
$3 \le x < 4$	16
$4 \le x < 5$	6
$5 \le x < 6$	5
$6 \le x < 7$	2

The following histogram displays the number of books on the x-axis and the frequency on the y-axis.



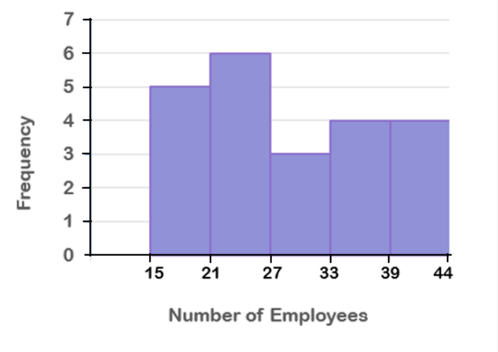
Note:
Try It
Exercise:

Problem:

The following data represent the number of employees at various restaurants in New York City. Using this data, create a histogram.

22351526 40281820 25343942 24221927 22344020 3828

Solution:



Note:
Count the money (bills and change) in your pocket or purse. Your instructor will record the amounts. As a class, construct a histogram displaying the data. Discuss how many intervals you think is appropriate. You may want to experiment with the number of intervals.

Frequency Polygons

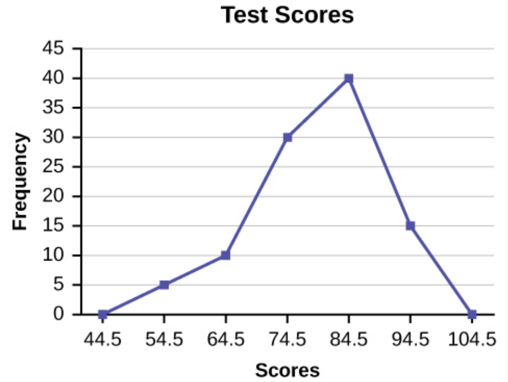
Frequency polygons are analogous to line graphs, and just as line graphs make continuous data visually easy to interpret, so too do frequency polygons.

To construct a frequency polygon, first examine the data and decide on the number of intervals, or class intervals, to use on the x-axis and y-axis. After choosing the appropriate ranges, begin plotting the data points. After all the points are plotted, draw line segments to connect them.

Example:
A frequency polygon was constructed from the frequency table below.

Frequency Distribution for Calculus Final Test Scores

Lower Bound	**Upper Bound**	**Frequency**	**Cumulative Frequency**
49.5	59.5	5	5
59.5	69.5	10	15
69.5	79.5	30	45
79.5	89.5	40	85
89.5	99.5	15	100



The first label on the x-axis is 44.5. This represents an interval extending from 39.5 to 49.5. Since the lowest test score is 54.5, this interval is used only to allow the graph to touch the x-axis. The point labeled 54.5 represents the next interval, or the first "real" interval from the table, and contains five scores. This reasoning is followed for each of the remaining intervals with the point 104.5 representing the interval from 99.5 to 109.5. Again, this interval contains no data and is only used so that the graph will touch the x-axis. Looking at the graph, we say that this distribution is skewed because one side of the graph does not mirror the other side.

Note:
Try It
Exercise:

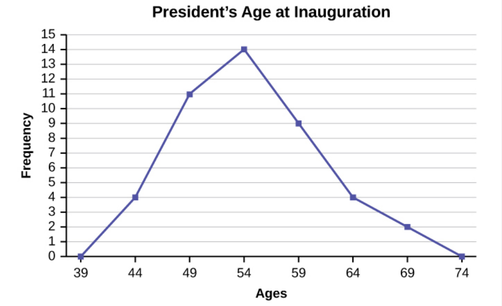
Problem: Construct a frequency polygon of U.S. Presidents' ages at inauguration shown in [link].

Age at Inauguration	Frequency
41.5–46.5	4
46.5–51.5	11
51.5–56.5	14
56.5–61.5	9
61.5–66.5	4
66.5–71.5	2

Solution:

The first label on the x-axis is 39. This represents an interval extending from 36.5 to 41.5. Since there are no ages less than 41.5, this interval is used only to allow the graph to touch the x-axis. The point labeled 44 represents the next interval, or the first "real" interval from the table, and contains four scores. This reasoning is followed for each of the remaining intervals with the point 74 representing the interval from 71.5 to 76.5. Again, this interval contains no data and is only used so that the graph will touch the x-axis.

Looking at the graph, we say that this distribution is skewed because one side of the graph does not mirror the other side.



Frequency polygons are useful for comparing distributions. This is achieved by overlaying the frequency polygons drawn for different data sets.

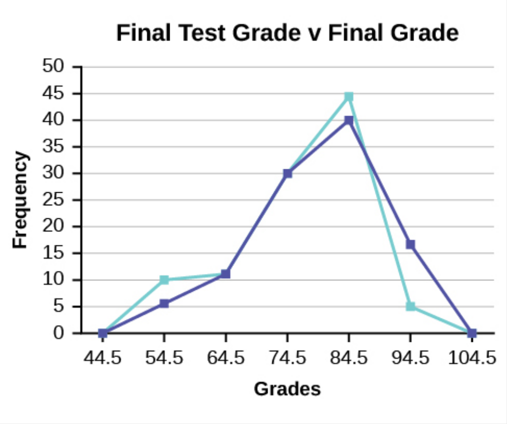
Example:

We will construct an overlay frequency polygon comparing the scores from [link] with the students' final numeric grade.

Frequency Distribution for Calculus Final Test Scores			
Lower Bound	**Upper Bound**	**Frequency**	**Cumulative Frequency**
49.5	59.5	5	5
59.5	69.5	10	15
69.5	79.5	30	45
79.5	89.5	40	85
89.5	99.5	15	100

Frequency Distribution for Calculus Final Grades			
Lower Bound	**Upper Bound**	**Frequency**	**Cumulative Frequency**
49.5	59.5	10	10
59.5	69.5	10	20

Frequency Distribution for Calculus Final Grades			
Lower Bound	**Upper Bound**	**Frequency**	**Cumulative Frequency**
69.5	79.5	30	50
79.5	89.5	45	95
89.5	99.5	5	100



Suppose that we want to study the temperature range of a region for an entire month. Every day at noon we note the temperature and write this down in a log. A variety of statistical studies could be done with this data. We could find the mean or the median temperature for the month. We could construct a histogram displaying the number of days that temperatures reach a certain range of values. However, all of these methods ignore a portion of the data that we have collected.

One feature of the data that we may want to consider is that of time. Since each date is paired with the temperature reading for the day, we don‘t have to think of the data as being random. We can instead use the times given to impose a chronological order on the data. A graph that recognizes this ordering and displays the changing temperature as the month progresses is called a time series graph.

Constructing a Time Series Graph

To construct a time series graph, we must look at both pieces of our **paired data set**. We start with a standard Cartesian coordinate system. The horizontal axis is used to plot the date or time increments, and the vertical axis is used to plot the values of the variable that we are measuring. By doing this, we make each point on the graph correspond to a date and a measured quantity. The points on the graph are typically connected by straight lines in the order in which they occur.

Example:
Exercise:

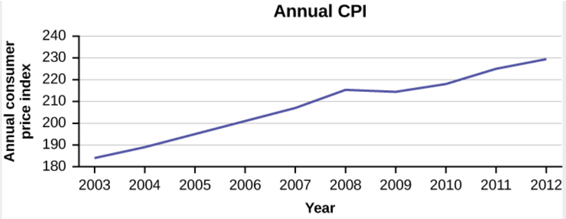
Problem:

The following data shows the Annual Consumer Price Index, each month, for ten years. Construct a time series graph for the Annual Consumer Price Index data only.

Year	**Jan**	**Feb**	**Mar**	**Apr**	**May**	**Jun**	**Jul**
2003	181.7	183.1	184.2	183.8	183.5	183.7	183.9
2004	185.2	186.2	187.4	188.0	189.1	189.7	189.4
2005	190.7	191.8	193.3	194.6	194.4	194.5	195.4
2006	198.3	198.7	199.8	201.5	202.5	202.9	203.5
2007	202.416	203.499	205.352	206.686	207.949	208.352	208.299
2008	211.080	211.693	213.528	214.823	216.632	218.815	219.964
2009	211.143	212.193	212.709	213.240	213.856	215.693	215.351
2010	216.687	216.741	217.631	218.009	218.178	217.965	218.011
2011	220.223	221.309	223.467	224.906	225.964	225.722	225.922
2012	226.665	227.663	229.392	230.085	229.815	229.478	229.104

Year	**Aug**	**Sep**	**Oct**	**Nov**	**Dec**	**Annual**
2003	184.6	185.2	185.0	184.5	184.3	184.0
2004	189.5	189.9	190.9	191.0	190.3	188.9
2005	196.4	198.8	199.2	197.6	196.8	195.3
2006	203.9	202.9	201.8	201.5	201.8	201.6
2007	207.917	208.490	208.936	210.177	210.036	207.342
2008	219.086	218.783	216.573	212.425	210.228	215.303
2009	215.834	215.969	216.177	216.330	215.949	214.537
2010	218.312	218.439	218.711	218.803	219.179	218.056
2011	226.545	226.889	226.421	226.230	225.672	224.939
2012	230.379	231.407	231.317	230.221	229.601	229.594

Solution:



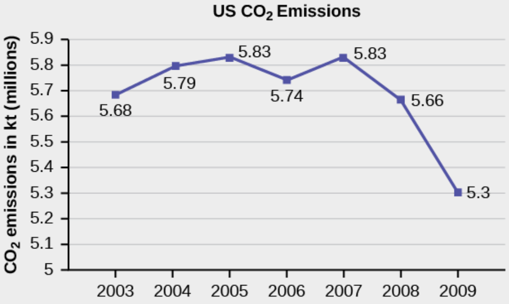
Note:
Try It
Exercise:

Problem:

The following table is a portion of a data set from www.worldbank.org. Use the table to construct a time series graph for CO_2 emissions for the United States.

CO2 Emissions			
	Ukraine	**United Kingdom**	**United States**
2003	352,259	540,640	5,681,664
2004	343,121	540,409	5,790,761
2005	339,029	541,990	5,826,394
2006	327,797	542,045	5,737,615
2007	328,357	528,631	5,828,697
2008	323,657	522,247	5,656,839
2009	272,176	474,579	5,299,563

Solution:



Uses of a Time Series Graph

Time series graphs are important tools in various applications of statistics. When recording values of the same variable over an extended period of time, sometimes it is difficult to discern any trend or pattern. However, once the same data points are displayed graphically, some features jump out. Time series graphs make trends easy to spot.

References

Data on annual homicides in Detroit, 1961–73, from Gunst & Mason’s book ‘Regression Analysis and its Application’, Marcel Dekker

“Timeline: Guide to the U.S. Presidents: Information on every president’s birthplace, political party, term of office, and more.” Scholastic, 2013. Available online at http://www.scholastic.com/teachers/article/timeline-guide-us-presidents (accessed April 3, 2013).

“Presidents.” Fact Monster. Pearson Education, 2007. Available online at http://www.factmonster.com/ipka/A0194030.html (accessed April 3, 2013).

“Food Security Statistics.” Food and Agriculture Organization of the United Nations. Available online at http://www.fao.org/economic/ess/ess-fs/en/ (accessed April 3, 2013).

“Consumer Price Index.” United States Department of Labor: Bureau of Labor Statistics. Available online at http://data.bls.gov/pdq/SurveyOutputServlet (accessed April 3, 2013).

“CO2 emissions (kt).” The World Bank, 2013. Available online at http://databank.worldbank.org/data/home.aspx (accessed April 3, 2013).

“Births Time Series Data.” General Register Office For Scotland, 2013. Available online at http://www.gro-scotland.gov.uk/statistics/theme/vital-events/births/time-series.html (accessed April 3, 2013).

“Demographics: Children under the age of 5 years underweight.” Indexmundi. Available online at http://www.indexmundi.com/g/r.aspx?t=50&v=2224&aml=en (accessed April 3, 2013).

Gunst, Richard, Robert Mason. *Regression Analysis and Its Application: A Data-Oriented Approach*. CRC Press: 1980.

“Overweight and Obesity: Adult Obesity Facts.” Centers for Disease Control and Prevention. Available online at http://www.cdc.gov/obesity/data/adult.html (accessed September 13, 2013).

Section Review

A **histogram** is a graphic version of a frequency distribution. The graph consists of bars of equal width drawn adjacent to each other. The horizontal scale represents class intervals of quantitative data values and the vertical scale represents frequencies. The heights of the bars correspond to frequency values. Histograms are typically used for large, continuous, quantitative data sets. A **frequency polygon** can also be used when graphing large data sets with data points that repeat. The data usually goes on x-axis with the frequency being graphed on the y-axis. **Time series graphs** can be helpful when looking at large amounts of data for one variable over a period of time.

Exercise:

Problem:

Sixty-five randomly selected car salespersons were asked the number of cars they generally sell in one week. Fourteen people answered that they generally sell three cars; nineteen generally sell four cars; twelve generally sell five cars; nine generally sell six cars; eleven generally sell seven cars. Complete the table.

Data Value (# cars)	Frequency	Relative Frequency	Cumulative Relative Frequency

Exercise:

Problem: What does the frequency column in [link] sum to? Why?

Solution:

65

Exercise:

Problem: What does the relative frequency column in [link] sum to? Why?

Exercise:

Problem: What is the difference between relative frequency and frequency for each data value in [link]?

Solution:

The relative frequency shows the *proportion* of data points that have each value. The frequency tells the *number* of data points that have each value.

Exercise:

Problem:

What is the difference between cumulative relative frequency and relative frequency for each data value?

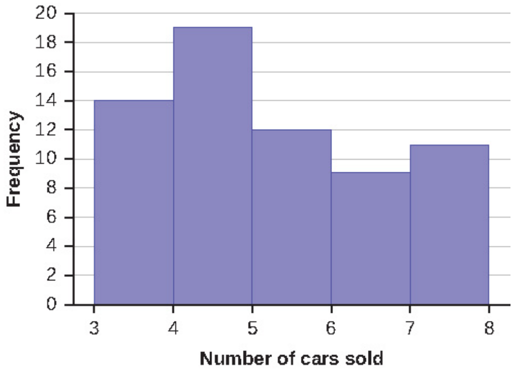
Exercise:

Problem:

To construct the histogram for the data in [link], determine appropriate minimum and maximum x and y values and the scaling. Sketch the histogram. Label the horizontal and vertical axes with words. Include numerical scaling.

Solution:

Answers will vary. One possible histogram is shown:



Exercise:

Problem: Construct a frequency polygon for the following:

a.

Pulse Rates for Women	Frequency
60–69	12
70–79	14
80–89	11
90–99	1
100–109	1
110–119	0
120–129	1

b.

Actual Speed in a 30 MPH Zone	Frequency
42–45	25
46–49	14
50–53	7
54–57	3
58–61	1

c.

Tar (mg) in Nonfiltered Cigarettes	Frequency
10–13	1
14–17	0
18–21	15
22–25	7
26–29	2

Exercise:

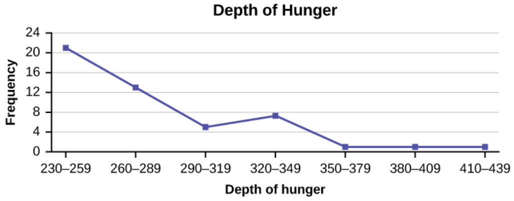
Problem:

Construct a frequency polygon from the frequency distribution for the 50 highest ranked countries for depth of hunger.

Depth of Hunger	Frequency
230–259	21
260–289	13
290–319	5
320–349	7
350–379	1
380–409	1
410–439	1

Solution:

Find the midpoint for each class. These will be graphed on the x-axis. The frequency values will be graphed on the y-axis values.



Exercise:

Problem:

Use the two frequency tables to compare the life expectancy of men and women from 20 randomly selected countries. Include an overlayed frequency polygon and discuss the shapes of the distributions, the center, the spread, and any outliers. What can we conclude about the life expectancy of women compared to men?

Life Expectancy at Birth – Women	**Frequency**
49–55	3
56–62	3
63–69	1
70–76	3
77–83	8
84–90	2

Life Expectancy at Birth – Men	**Frequency**
49–55	3
56–62	3
63–69	1
70–76	1
77–83	7
84–90	5

Exercise:

Problem:

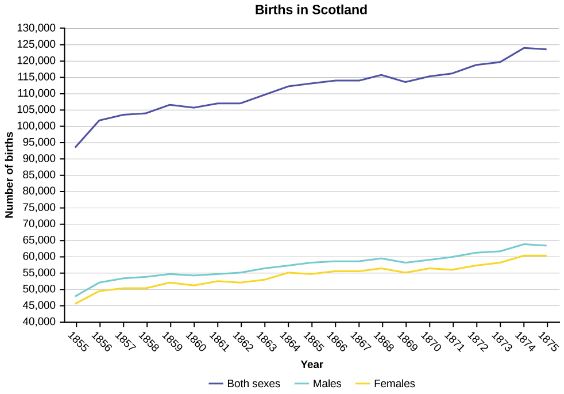
Construct a times series graph for (a) the number of male births, (b) the number of female births, and (c) the total number of births.

Sex/Year	1855	1856	1857	1858	1859	1860	1861
Female	45,545	49,582	50,257	50,324	51,915	51,220	52,403
Male	47,804	52,239	53,158	53,694	54,628	54,409	54,606
Total	93,349	101,821	103,415	104,018	106,543	105,629	107,009

Sex/Year	1862	1863	1864	1865	1866	1867	1868	18
Female	51,812	53,115	54,959	54,850	55,307	55,527	56,292	55
Male	55,257	56,226	57,374	58,220	58,360	58,517	59,222	58
Total	107,069	109,341	112,333	113,070	113,667	114,044	115,514	11

Sex/Year	1871	1870	1872	1871	1872	1827	1874	18
Female	56,099	56,431	57,472	56,099	57,472	58,233	60,109	6
Male	60,029	58,959	61,293	60,029	61,293	61,467	63,602	6
Total	116,128	115,390	118,765	116,128	118,765	119,700	123,711	1

Solution:



Exercise:

Problem:

The following data sets list full time police per 100,000 citizens along with homicides per 100,000 citizens for the city of Detroit, Michigan during the period from 1961 to 1973.

Year	1961	1962	1963	1964	1965	1966	1967
Police	260.35	269.8	272.04	272.96	272.51	261.34	268.89
Homicides	8.6	8.9	8.52	8.89	13.07	14.57	21.36

Year	1968	1969	1970	1971	1972	1973
Police	295.99	319.87	341.43	356.59	376.69	390.19
Homicides	28.03	31.49	37.39	46.26	47.24	52.33

a. Construct a double time series graph using a common x-axis for both sets of data.
b. Which variable increased the fastest? Explain.
c. Did Detroit’s increase in police officers have an impact on the murder rate? Explain.

Homework

Exercise:

Problem:

Suppose that three book publishers were interested in the number of fiction paperbacks adult consumers purchase per month. Each publisher conducted a survey. In the survey, adult consumers were asked the number of fiction paperbacks they had purchased the previous month. The results are as follows:

# of books	Freq.	Rel. Freq.
0	10	
1	12	
2	16	
3	12	
4	8	
5	6	
6	2	
8	2	

Publisher A

# of books	Freq.	Rel. Freq.
0	18	
1	24	
2	24	
3	22	
4	15	
5	10	
7	5	
9	1	

Publisher B

# of books	Freq.	Rel. Freq.
0–1	20	
2–3	35	
4–5	12	
6–7	2	
8–9	1	

Publisher C

a. Find the relative frequencies for each survey. Write them in the charts.
b. Using either a graphing calculator, computer, or by hand, use the frequency column to construct a histogram for each publisher's survey. For Publishers A and B, make bar widths of one. For Publisher C, make bar widths of two.
c. In complete sentences, give two reasons why the graphs for Publishers A and B are not identical.
d. Would you have expected the graph for Publisher C to look like the other two graphs? Why or why not?
e. Make new histograms for Publisher A and Publisher B. This time, make bar widths of two.
f. Now, compare the graph for Publisher C to the new graphs for Publishers A and B. Are the graphs more similar or more different? Explain your answer.

Exercise:

Problem:

Often, cruise ships conduct all on-board transactions, with the exception of gambling, on a cashless basis. At the end of the cruise, guests pay one bill that covers all onboard transactions. Suppose that 60 single travelers and 70 couples were surveyed as to their on-board bills for a seven-day cruise from Los Angeles to the Mexican Riviera. Following is a summary of the bills for each group.

Amount($)	**Frequency**	**Rel. Frequency**
50–99	5	
100–149	10	
150–199	15	
200–249	15	
250–199	10	
300–349	5	

Singles

Amount($)	**Frequency**	**Rel. Frequency**
100–149	5	
200–249	5	
250–299	5	
300–349	5	
350–399	10	
400–449	10	
450–499	10	
500–549	10	
550–599	5	
600–649	5	

Couples

a. Fill in the relative frequency for each group.
b. Construct a histogram for the singles group. Scale the x-axis by \$50 widths. Use relative frequency on the y-axis.
c. Construct a histogram for the couples group. Scale the x-axis by \$50 widths. Use relative frequency on the y-axis.
d. Compare the two graphs:
    i. List two similarities between the graphs.
    ii. List two differences between the graphs.
    iii. Overall, are the graphs more similar or different?
e. Construct a new graph for the couples by hand. Since each couple is paying for two individuals, instead of scaling the x-axis by \$50, scale it by \$100. Use relative frequency on the y-axis.
f. Compare the graph for the singles with the new graph for the couples:
    i. List two similarities between the graphs.
    ii. Overall, are the graphs more similar or different?
g. How did scaling the couples graph differently change the way you compared it to the singles graph?
h. Based on the graphs, do you think that individuals spend the same amount, more or less, as singles as they do person by person as a couple? Explain why in one or two complete sentences.

Solution:

Amount($)	**Frequency**	**Relative Frequency**
51–100	5	0.08
101–150	10	0.17
151–200	15	0.25
201–250	15	0.25
251–300	10	0.17
301–350	5	0.08

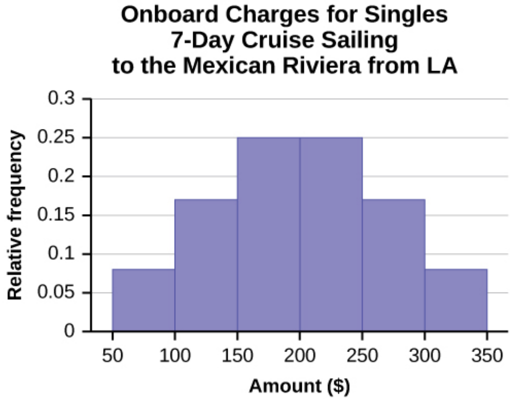
Singles

Amount($)	**Frequency**	**Relative Frequency**
100–150	5	0.07

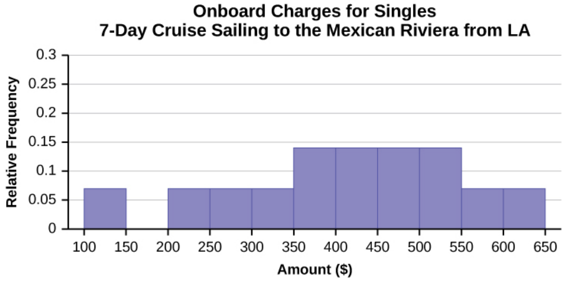
Amount($)	Frequency	Relative Frequency
201–250	5	0.07
251–300	5	0.07
301–350	5	0.07
351–400	10	0.14
401–450	10	0.14
451–500	10	0.14
501–550	10	0.14
551–600	5	0.07
601–650	5	0.07

Couples

a. See [link] and [link].
b. When reading the histogram, recall that data values that fall on the left boundary of a class interval are included, while data values that fall on the right boundary are not included in that class interval.



c. In the following histogram, the data values that fall on the right boundary are counted in the class interval, while values that fall on the left boundary are not counted (with the exception of the first interval where values on both boundaries are included).



d. Compare the two graphs:

   i. Answers may vary. Possible answers include:

      - Both graphs have a single peak.

        - Both graphs use class intervals with width equal to $50.

    ii. Answers may vary. Possible answers include:

        - The couples graph has a class interval with no values.
        - It takes almost twice as many class intervals to display the data for couples.

    iii. Answers may vary. Possible answers include: The graphs are more similar than different because the overall patterns for the graphs are the same.

e. Check student's solution.
f. Compare the graph for the Singles with the new graph for the Couples:

    i.
        - Both graphs have a single peak.
        - Both graphs display 6 class intervals.
        - Both graphs show the same general pattern.

    ii. Answers may vary. Possible answers include: Although the width of the class intervals for couples is double that of the class intervals for singles, the graphs are more similar than they are different.

g. Answers may vary. Possible answers include: You are able to compare the graphs interval by interval. It is easier to compare the overall patterns with the new scale on the Couples graph. Because a couple represents two individuals, the new scale leads to a more accurate comparison.
h. Answers may vary. Possible answers include: Based on the histograms, it seems that spending does not vary much from singles to individuals who are part of a couple. The overall patterns are the same. The range of spending for couples is approximately double the range for individuals.

Exercise:

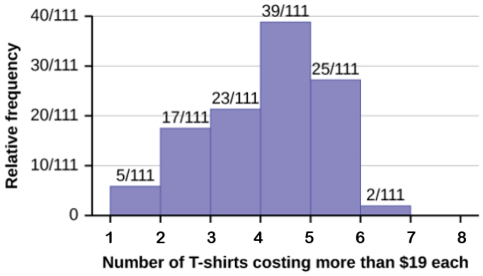
Problem:

Twenty-five randomly selected students were asked the number of movies they watched the previous week. The results are as follows.

# of movies	Frequency	Relative Frequency	Cumulative Relative Frequency
0	5		
1	9		
2	6		
3	4		
4	1		

a. Construct a histogram of the data.
b. Complete the columns of the chart.

Use the following information to answer the next two exercises: Suppose one hundred eleven people who shopped in a special t-shirt store were asked the number of t-shirts they own costing more than $19 each.



Exercise:

Problem:

The percentage of people who own at most three t-shirts costing more than $19 each is approximately:

a. 21
b. 59
c. 41
d. Cannot be determined

Solution:

c

Exercise:

Problem:

If the data were collected by asking the first 111 people who entered the store, then the type of sampling is:

a. cluster
b. simple random
c. stratified
d. convenience

Glossary

Frequency
the number of times a value of the data occurs

Histogram
a graphical representation in x-y form of the distribution of data in a data set; x represents the data and y represents the frequency, or relative frequency. The graph consists of contiguous rectangles.

Relative Frequency
the ratio of the number of times a value of the data occurs in the set of all outcomes to the number of all outcomes

Measures of the Location of the Data

The common measures of location are **quartiles** and **percentiles**

Quartiles are special percentiles. The first quartile, Q_1, is the same as the 25th percentile, and the third quartile, Q_3, is the same as the 75th percentile. The median, M, is called both the second quartile and the 50th percentile.

To calculate quartiles and percentiles, the data must be ordered from smallest to largest. Quartiles divide ordered data into quarters. Percentiles divide ordered data into hundredths. To score in the 90th percentile of an exam does not mean, necessarily, that you received 90% on a test. It means that 90% of test scores are the same or less than your score and 10% of the test scores are the same or greater than your test score.

Percentiles are useful for comparing values. For this reason, universities and colleges use percentiles extensively. One instance in which colleges and universities use percentiles is when SAT results are used to determine a minimum testing score that will be used as an acceptance factor. For example, suppose Duke accepts SAT scores at or above the 75th percentile. That translates into a score of at least 1220.

Percentiles are mostly used with very large populations. Therefore, if you were to say that 90% of the test scores are less (and not the same or less) than your score, it would be acceptable because removing one particular data value is not significant.

The **median** is a number that measures the "center" of the data. You can think of the median as the "middle value," but it does not actually have to be one of the observed values. It is a number that separates ordered data into halves. Half the values are the same number or smaller than the median, and half the values are the same number or larger. For example, consider the following data.
1; 11.5; 6; 7.2; 4; 8; 9; 10; 6.8; 8.3; 2; 2; 10; 1
Ordered from smallest to largest:
1; 1; 2; 2; 4; 6; 6.8; 7.2; 8; 8.3; 9; 10; 10; 11.5

Since there are 14 observations, the median is between the seventh value, 6.8, and the eighth value, 7.2. To find the median, add the two values together and divide by two.

Equation:

$$\frac{6.8+7.2}{2}=7$$

The median is seven. Half of the values are smaller than seven and half of the values are larger than seven.

Quartiles are numbers that separate the data into quarters. Quartiles may or may not be part of the data. To find the quartiles, first find the median or second quartile. The first quartile, Q_1, is the middle value of the lower half of the data, and the third quartile, Q_3, is the middle value, or median, of the upper half of the data. To get the idea, consider the same data set:
1; 1; 2; 2; 4; 6; 6.8; 7.2; 8; 8.3; 9; 10; 10; 11.5

The median or **second quartile** is seven. The lower half of the data are 1, 1, 2, 2, 4, 6, 6.8. The middle value of the lower half is two.
1; 1; 2; 2; 4; 6; 6.8

The number two, which is part of the data, is the **first quartile**. One-fourth of the entire sets of values are the same as or less than two and three-fourths of the values are more than two.

The upper half of the data is 7.2, 8, 8.3, 9, 10, 10, 11.5. The middle value of the upper half is nine.

The **third quartile**, Q_3, is nine. Three-fourths (75%) of the ordered data set are less than nine. One-fourth (25%) of the ordered data set are greater than nine. The third quartile is part of the data set in this example.

The **interquartile range** is a number that indicates the spread of the middle half or the middle 50% of the data. It is the difference between the third quartile (Q_3) and the first quartile (Q_1).

$IQR = Q_3 – Q_1$

The *IQR* can help to determine potential **outliers**. **A value is suspected to be a potential outlier if it is less than (1.5)(*IQR*) below the first quartile or more than (1.5)(*IQR*) above the third quartile**. Potential outliers always require further investigation.

Note:
NOTE
A potential outlier is a data point that is significantly different from the other data points. These special data points may be errors or some kind of abnormality or they may be a key to understanding the data.

Example:
Exercise:

Problem:

For the following 13 real estate prices, calculate the *IQR* and determine if any prices are potential outliers. Prices are in dollars.
389,950; 230,500; 158,000; 479,000; 639,000; 114,950; 5,500,000; 387,000; 659,000; 529,000; 575,000; 488,800; 1,095,000

Solution:

Order the data from smallest to largest.
114,950; 158,000; 230,500; 387,000; 389,950; 479,000; 488,800; 529,000; 575,000; 639,000; 659,000; 1,095,000; 5,500,000

$M = 488,800$

$Q_1 = \frac{230{,}500 + 387{,}000}{2} = 308{,}750$

$Q_3 = \frac{639{,}000 + 659{,}000}{2} = 649{,}000$

IQR = 649,000 – 308,750 = 340,250

(1.5)(IQR) = (1.5)(340,250) = 510,375

Q_1 – (1.5)(IQR) = 308,750 – 510,375 = –201,625

Q_3 + (1.5)(IQR) = 649,000 + 510,375 = 1,159,375

No house price is less than –201,625. However, 5,500,000 is more than 1,159,375. Therefore, 5,500,000 is a potential **outlier**.

Note:
Try It
Exercise:

Problem:

For the following 11 salaries, calculate the *IQR* and determine if any salaries are outliers. The salaries are in dollars.

$33,000 $64,500 $28,000 $54,000 $72,000 $68,500 $69,000 $42,000 $54,000 $120,000 $40,500

Solution:

Order the data from smallest to largest.

$28,000 $33,000 $40,500 $42,000 $54,000 $54,000 $64,500 $68,500 $69,000 $72,000 $120,000

Median = $54,000

Q_1 = $40,500

Q_3 = $69,000

$IQR = \$69{,}000 – \$40{,}500 = \$28{,}500$

$(1.5)(IQR) = (1.5)(\$28{,}500) = \$42{,}750$

$Q_1 – (1.5)(IQR) = \$40{,}500 – \$42{,}750 = –\$2{,}250$

$Q_3 + (1.5)(IQR) = \$69{,}000 + \$42{,}750 = \$111{,}750$

No salary is less than –\$2,250. However, \$120,000 is more than \$11,750, so \$120,000 is a potential outlier.

Example:
Exercise:

Problem:

For the two data sets in the test scores example, find the following:

a. The interquartile range. Compare the two interquartile ranges.
b. Any outliers in either set.

Solution:

The five number summary for the day and night classes is

	Minimum	Q_1	**Median**	Q_3	**Maximum**
Day	32	56	74.5	82.5	99
Night	25.5	78	81	89	98

a. The IQR for the day group is $Q_3 – Q_1 = 82.5 – 56 = 26.5$

The IQR for the night group is $Q_3 – Q_1 = 89 – 78 = 11$

The interquartile range (the spread or variability) for the day class is larger than the night class *IQR*. This suggests more variation will be found in the day class's class test scores.

b. Day class outliers are found using the IQR times 1.5 rule. So,

    - $Q_1 - IQR(1.5) = 56 – 26.5(1.5) = 16.25$
    - $Q_3 + IQR(1.5) = 82.5 + 26.5(1.5) = 122.25$

   Since the minimum and maximum values for the day class are greater than 16.25 and less than 122.25, there are no outliers.

   Night class outliers are calculated as:

    - $Q_1 – IQR\ (1.5) = 78 – 11(1.5) = 61.5$
    - $Q_3 + \mathrm{IQR}(1.5) = 89 + 11(1.5) = 105.5$

   For this class, any test score less than 61.5 is an outlier. Therefore, the scores of 45 and 25.5 are outliers. Since no test score is greater than 105.5, there is no upper end outlier.

Note:
Try It
Exercise:

Problem:

Find the interquartile range for the following two data sets and compare them.

Test Scores for Class *A*
69; 96; 81; 79; 65; 76; 83; 99; 89; 67; 90; 77; 85; 98; 66; 91; 77; 69; 80; 94
Test Scores for Class *B*
90; 72; 80; 92; 90; 97; 92; 75; 79; 68; 70; 80; 99; 95; 78; 73; 71; 68; 95; 100

Solution:

Class *A*

Order the data from smallest to largest.

65 66 67 69 69 76 77 77 79 80 81 83 85 89 90 91 94 96 98 99

$\text{Median} = \frac{80+81}{2} = 80.5$

$Q_1 = \frac{69+76}{2} = 72.5$

$Q_3 = \frac{90+91}{2} = 90.5$

IQR = 90.5 – 72.5 = 18

Class *B*

Order the data from smallest to largest.

68 68 70 71 72 73 75 78 79 80 80 90 90 92 92 95 95 97 99 100

$\text{Median} = \frac{80+80}{2} = 80$

$Q_1 = \frac{72+73}{2} = 72.5$

$Q_3 = \frac{92+95}{2} = 93.5$

IQR = 93.5 – 72.5 = 21

The data for Class *B* has a larger *IQR*, so the scores between Q_3 and Q_1 (middle 50%) for the data for Class *B* are more spread out and not clustered about the median.

Example:
Fifty statistics students were asked how much sleep they get per school night (rounded to the nearest hour). The results were:

AMOUNT OF SLEEP PER SCHOOL NIGHT (HOURS)	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
4	2	0.04	0.04
5	5	0.10	0.14
6	7	0.14	0.28
7	12	0.24	0.52
8	14	0.28	0.80
9	7	0.14	0.94
10	3	0.06	1.00

Find the 28th percentile. Notice the 0.28 in the "cumulative relative frequency" column. Twenty-eight percent of 50 data values is 14 values. There are 14 values less than the 28th percentile. They include the two 4s, the five 5s, and the seven 6s. The 28th percentile is between the last six and the first seven. **The 28th percentile is 6.5.**

Find the median. Look again at the "cumulative relative frequency" column and find 0.52. The median is the 50th percentile or the second quartile. 50% of 50 is 25. There are 25 values less than the median. They include the two 4s, the five 5s, the seven 6s, and eleven of the 7s. The median or 50th percentile is between the 25th, or seven, and 26th, or seven, values. **The median is seven.**

Find the third quartile. The third quartile is the same as the 75th percentile. You can "eyeball" this answer. If you look at the "cumulative relative frequency" column, you find 0.52 and 0.80. When you have all the fours, fives, sixes and sevens, you have 52% of the data. When you include all the eights, you have 80% of the data. **The 75th percentile, then, must be an eight**. Another way to look at the problem is to find 75% of 50, which is 37.5, and round up to 38. The third quartile, Q_3, is the 38th value, which is an eight. You can check this answer by counting the values. (There are 37 values below the third quartile and 12 values above.)

Note:
Try it
Exercise:

Problem:

Forty bus drivers were asked how many hours they spend each day running their routes (rounded to the nearest hour). Find the 65th percentile.

Amount of time spent on route (hours)	Frequency	Relative Frequency	Cumulative Relative Frequency
2	12	0.30	0.30
3	14	0.35	0.65
4	10	0.25	0.90
5	4	0.10	1.00

Solution:

The 65th percentile is between the last three and the first four.

The 65th percentile is 3.5.

Example:
Exercise:

Problem: Using [link]:

a. Find the 80th percentile.
b. Find the 90th percentile.
c. Find the first quartile. What is another name for the first quartile?

Solution:

Using the data from the frequency table, we have:

a. The 80th percentile is between the last eight and the first nine in the table (between the 40th and 41st values). Therefore, we need to take the mean of the 40th an 41st values. The 80th percentile $= \frac{8+9}{2} = 8.5$
b. The 90th percentile will be the 45th data value (location is 0.90(50) = 45) and the 45th data value is nine.
c. Q_1 is also the 25th percentile. The 25th percentile location calculation: P_{25} = 0.25(50) = 12.5 ≈ 13 the 13th data value. Thus, the 25th percentile is six.

Note:
Try It
Exercise:

Problem:

Refer to the [link]. Find the third quartile. What is another name for the third quartile?

Solution:

The third quartile is the 75th percentile, which is four. The 65th percentile is between three and four, and the 90th percentile is between four and 5.75. The third quartile is between 65 and 90, so it must be four.

Note:
Collaborative Statistics
Your instructor or a member of the class will ask everyone in class how many sweaters they own. Answer the following questions:

1. How many students were surveyed?
2. What kind of sampling did you do?
3. Construct two different histograms. For each, starting value = _____ ending value = ____.
4. Find the median, first quartile, and third quartile.
5. Construct a table of the data to find the following:
    a. the 10^{th} percentile
    b. the 70^{th} percentile
    c. the percent of students who own less than four sweaters

A Formula for Finding the *k*th Percentile

If you were to do a little research, you would find several formulas for calculating the k^{th} percentile. Here is one of them.

k = the k^{th} percentile. It may or may not be part of the data.

i = the index (ranking or position of a data value)

n = the total number of data values

- Order the data from smallest to largest.
- Calculate $i = \frac{k}{100}(n+1)$
- If i is an integer, then the k^{th} percentile is the data value in the i^{th} position in the ordered set of data.
- If i is not an integer, then round i up and round i down to the nearest integers. Average the two data values in these two positions in the ordered data set. This is easier to understand in an example.

Example:

Exercise:

Problem:

Listed are 29 ages for Academy Award winning best actors *in order from smallest to largest.*
18; 21; 22; 25; 26; 27; 29; 30; 31; 33; 36; 37; 41; 42; 47; 52; 55; 57; 58; 62; 64; 67; 69; 71; 72; 73; 74; 76; 77

a. Find the 70th percentile.
b. Find the 83rd percentile.

Solution:

a.
   - k = 70
   - i = the index
   - n = 29

$i = \frac{k}{100}(n + 1) = (\frac{70}{100})(29 + 1) = 21$. Twenty-one is an integer, and the data value in the 21st position in the ordered data set is 64. The 70th percentile is 64 years.

b.
   - k = 83rd percentile
   - i = the index
   - n = 29

$i = \frac{k}{100}(n + 1) =)\frac{83}{100})(29 + 1) = 24.9$, which is NOT an integer. Round it down to 24 and up to 25. The age in the 24th position is 71 and the age in the 25th position is 72. Average 71 and 72. The 83rd percentile is 71.5 years.

Note:
Try It
Exercise:

Problem:

Listed are 29 ages for Academy Award winning best actors *in order from smallest to largest.*

18; 21; 22; 25; 26; 27; 29; 30; 31; 33; 36; 37; 41; 42; 47; 52; 55; 57; 58; 62; 64; 67; 69; 71; 72; 73; 74; 76; 77
Calculate the 20^{th} percentile and the 55^{th} percentile.

Solution:

k = 20. Index = i = $\frac{k}{100}(n+1) = \frac{20}{100}$(29 + 1) = 6. The age in the sixth position is 27. The 20^{th} percentile is 27 years.

k = 55. Index = i = $\frac{k}{100}(n+1) = \frac{55}{100}$(29 + 1) = 16.5. Round down to 16 and up to 17. The age in the 16^{th} position is 52 and the age in the 17^{th} position is 55. The average of 52 and 55 is 53.5. The 55^{th} percentile is 53.5 years.

Note:
NOTE
You can calculate percentiles using calculators and computers. There are a variety of online calculators.

A Formula for Finding the Percentile of a Value in a Data Set

Suppose you took a test and you want to know what percentile your test score is. If you have a list of the test scores for all the students who took the test, you can use the following method to determine your percentile.

Note: This is quite different than finding the k^{th} percentile, as we saw earlier.

- Order the data from smallest to largest.
- x = the number of data values counting from the bottom of the data list up to, but not including, the data value for which you want to find the

percentile.
- y = the number of data values equal to the particular data value for which you want to find the percentile.
- n = the total number of data values.
- Calculate $k = \frac{x+0.5y}{n}(100)$. Then round k to the nearest integer.

Example:
Exercise:

Problem:

Listed are 29 ages for Academy Award winning best actors *in order from smallest to largest.*
18; 21; 22; 25; 26; 27; 29; 30; 31; 33; 36; 37; 41; 42; 47; 52; 55; 57; 58; 62; 64; 67; 69; 71; 72; 73; 74; 76; 77

a. Find the percentile for 58.
b. Find the percentile for 25.

Solution:

a. Counting from the bottom of the list, there are 18 data values less than 58. There is one value of 58.

   x = 18 and y = 1.
   $k = \frac{x+0.5y}{n}(100) = \frac{18+0.5(1)}{29}(100) = 63.80$, which rounds to 64.
   Therefore, 58 is the 64th percentile.
b. Counting from the bottom of the list, there are three data values less than 25. There is one value of 25.

   x = 3 and y = 1.
   $k = \frac{x+0.5y}{n}(100) = \frac{3+0.5(1)}{29}(100) = 12.07$, which rounds to 12.
   Therefore, twenty-five is the 12th percentile.

Note:
Try It
Exercise:

Problem:

Listed are 30 ages for Academy Award winning best actors in order from smallest to largest.

18; 21; 22; 25; 26; 27; 29; 30; 31, 31; 33; 36; 37; 41; 42; 47; 52; 55; 57; 58; 62; 64; 67; 69; 71; 72; 73; 74; 76; 77
Find the percentiles for 47 and 31.

Solution:

Percentile for 47: Counting from the bottom of the list, there are 15 data values less than 47. There is one value of 47.

x = 15 and y = 1.
$k = \frac{x+0.5y}{n}(100) = \frac{15+0.5(1)}{30}(100) = 51.67$, which rounds to 52.
Therefore, 47 is the 52nd percentile.

Percentile for 31: Counting from the bottom of the list, there are eight data values less than 31. There are two values of 31.

x = 8 and y = 2.
$k = \frac{x+0.5y}{n}(100) = \frac{8+0.5(2)}{30}(100) = 30$.
Therefore, 31 is the 30th percentile.

Interpreting Percentiles, Quartiles, and Median

A percentile indicates the relative standing of a data value when data are sorted into numerical order from smallest to largest. Percentages of data values are less than or equal to the kth percentile. For example, 15% of data values are less than or equal to the 15th percentile.

- Low percentiles always correspond to lower data values.

- High percentiles always correspond to higher data values.

A percentile may or may not correspond to a value judgment about whether it is "good" or "bad." The interpretation of whether a certain percentile is "good" or "bad" depends on the context of the situation to which the data applies. In some situations, a low percentile would be considered "good;" in other contexts a high percentile might be considered "good". In many situations, there is no value judgment that applies.

Understanding how to interpret percentiles properly is important not only when describing data, but also when calculating probabilities in later chapters of this text.

Note:
Guideline
When writing the interpretation of a percentile in the context of the given data, the sentence should contain the following information.

- information about the context of the situation being considered
- the data value (value of the variable) that represents the percentile
- the percent of individuals or items with data values below the percentile
- the percent of individuals or items with data values above the percentile.

Example:
Exercise:

Problem:

On a timed math test, the first quartile for time it took to finish the exam was 35 minutes. Interpret the first quartile in the context of this situation.

Solution:

- Twenty-five percent of students finished the exam in 35 minutes or less.

- Seventy-five percent of students finished the exam in 35 minutes or more.
- A low percentile could be considered good, as finishing more quickly on a timed exam is desirable. (If you take too long, you might not be able to finish.)

Note:
Try It
Exercise:

Problem:

For the 100-meter dash, the third quartile for times for finishing the race was 11.5 seconds. Interpret the third quartile in the context of the situation.

Solution:

Twenty-five percent of runners finished the race in 11.5 seconds or more. Seventy-five percent of runners finished the race in 11.5 seconds or less. A lower percentile is good because finishing a race more quickly is desirable.

Example:
Exercise:

Problem:

On a 20 question math test, the 70th percentile for number of correct answers was 16. Interpret the 70th percentile in the context of this situation.

Solution:

- Seventy percent of students answered 16 or fewer questions correctly.
- Thirty percent of students answered 16 or more questions correctly.
- A higher percentile could be considered good, as answering more questions correctly is desirable.

Note:
Try It
Exercise:

Problem:

On a 60 point written assignment, the 80th percentile for the number of points earned was 49. Interpret the 80th percentile in the context of this situation.

Solution:

Eighty percent of students earned 49 points or fewer. Twenty percent of students earned 49 or more points. A higher percentile is good because getting more points on an assignment is desirable.

Example:
Exercise:

Problem:

At a community college, it was found that the 30th percentile of credit units that students are enrolled for is seven units. Interpret the 30th percentile in the context of this situation.

Solution:

- Thirty percent of students are enrolled in seven or fewer credit units.
- Seventy percent of students are enrolled in seven or more credit units.
- In this example, there is no "good" or "bad" value judgment associated with a higher or lower percentile. Students attend community college for varied reasons and needs, and their course load varies according to their needs.

Note:
Try It

Exercise:

Problem:

During a season, the 40th percentile for points scored per player in a game is eight. Interpret the 40th percentile in the context of this situation.

Solution:

Forty percent of players scored eight points or fewer. Sixty percent of players scored eight points or more. A higher percentile is good because getting more points in a basketball game is desirable.

Example:

Sharpe Middle School is applying for a grant that will be used to add fitness equipment to the gym. The principal surveyed 15 anonymous students to determine how many minutes a day the students spend exercising. The results from the 15 anonymous students are shown.

0 minutes; 40 minutes; 60 minutes; 30 minutes; 60 minutes
10 minutes; 45 minutes; 30 minutes; 300 minutes; 90 minutes;
30 minutes; 120 minutes; 60 minutes; 0 minutes; 20 minutes

Determine the following five values.

- Min = 0
- Q_1 = 20
- Med = 40
- Q_3 = 60
- Max = 300

If you were the principal, would you be justified in purchasing new fitness equipment? Since 75% of the students exercise for 60 minutes or less daily, and since the *IQR* is 40 minutes (60 – 20 = 40), we know that half of the students surveyed exercise between 20 minutes and 60 minutes daily. This seems a reasonable amount of time spent exercising, so the principal would be justified in purchasing the new equipment.

However, the principal needs to be careful. The value 300 appears to be a potential outlier.

$Q_3 + 1.5(IQR) = 60 + (1.5)(40) = 120$.

The value 300 is greater than 120 so it is a potential outlier. If we delete it and calculate the five values, we get the following values:

- Min = 0
- Q_1 = 20
- Q_3 = 60
- Max = 120

We still have 75% of the students exercising for 60 minutes or less daily and half of the students exercising between 20 and 60 minutes a day. However, 15 students is a small sample and the principal should survey more students to be sure of his survey results.

References

Cauchon, Dennis, Paul Overberg. “Census data shows minorities now a majority of U.S. births.” USA Today, 2012. Available online at http://usatoday30.usatoday.com/news/nation/story/2012-05-17/minority-birthscensus/55029100/1 (accessed April 3, 2013).

Data from the United States Department of Commerce: United States Census Bureau. Available online at http://www.census.gov/ (accessed April 3, 2013).

“1990 Census.” United States Department of Commerce: United States Census Bureau. Available online at http://www.census.gov/main/www/cen1990.html (accessed April 3, 2013).

Data from *San Jose Mercury News*.

Data from *Time Magazine*; survey by Yankelovich Partners, Inc.

Section Review

The values that divide a rank-ordered set of data into 100 equal parts are called percentiles. Percentiles are used to compare and interpret data. For example, an observation at the 50th percentile would be greater than 50 percent of the other obeservations in the set. Quartiles divide data into quarters. The first quartile (Q_1) is the 25th percentile,the second quartile (Q_2 or median) is 50th percentile, and the third quartile (Q_3) is the the 75$^{\text{th}}$ percentile. The interquartile range, or *IQR*, is the range of the middle 50 percent of the data values. The *IQR* is found by subtracting Q_1 from Q_3, and can help determine outliers by using the following two expressions.

- $Q_3 + IQR(1.5)$
- $Q_1 - IQR(1.5)$

Formula Review

$i = \left(\frac{k}{100}\right)(n+1)$

where i = the ranking or position of a data value,

k = the k^{th} percentile,

n = total number of data.

Formula for finding the percentile of a data value:
$k = \left(\frac{x + 0.5y}{n}\right)(100)$

where k is always rounded to the nearest integer,

x = the number of values counting from the bottom of the data list up to but not including the particular data value for which you want to find the percentile,

y = the number of data values equal to the particular data value for which you want to find the percentile,

and n = total number of data values.

Exercise:

Problem:

Listed are 29 ages for Academy Award winning best actors *in order from smallest to largest.*

18; 21; 22; 25; 26; 27; 29; 30; 31; 33; 36; 37; 41; 42; 47; 52; 55; 57; 58; 62; 64; 67; 69; 71; 72; 73; 74; 76; 77

a. Find the 40th percentile.
b. Find the 78th percentile.

Solution:

a. The 40th percentile is 37 years.
b. The 78th percentile is 70 years.

Exercise:

Problem:

Listed are 32 ages for Academy Award winning best actors *in order from smallest to largest.*

18; 18; 21; 22; 25; 26; 27; 29; 30; 31; 31; 33; 36; 37; 37; 41; 42; 47; 52; 55; 57; 58; 62; 64; 67; 69; 71; 72; 73; 74; 76; 77

a. Find the percentile of 37.
b. Find the percentile of 72.

Exercise:

Problem:

Jesse was ranked 37th in his graduating class of 180 students. At what percentile is Jesse's ranking?

Solution:

Jesse graduated 37th out of a class of 180 students. There are 180 – 37 = 143 students ranked below Jesse. There is one rank of 37.

x = 143 and y = 1.
$k = \frac{x+0.5y}{n}$(100) = $\frac{143+0.5(1)}{180}$(100) = 79.72, which rounds to 80.
Jesse's rank of 37 puts him at the 80th percentile.

Exercise:

Problem:

a. For runners in a race, a low time means a faster run. The winners in a race have the shortest running times. Is it more desirable to have a finish time with a high or a low percentile when running a race?
b. The 20th percentile of run times in a particular race is 5.2 minutes. Write a sentence interpreting the 20th percentile in the context of the situation.
c. A bicyclist in the 90th percentile of a bicycle race completed the race in 1 hour and 12 minutes. Is he among the fastest or slowest cyclists in the race? Write a sentence interpreting the 90th percentile in the context of the situation.

Exercise:

Problem:

a. For runners in a race, a higher speed means a faster run. Is it more desirable to have a speed with a high or a low percentile when running a race?
b. The 40th percentile of speeds in a particular race is 7.5 miles per hour. Write a sentence interpreting the 40th percentile in the context of the situation.

Solution:

a. For runners in a race it is more desirable to have a high percentile for speed. A high percentile means a higher speed which is faster.
b. 40% of runners ran at speeds of 7.5 miles per hour or less (slower). 60% of runners ran at speeds of 7.5 miles per hour or more (faster).

Exercise:

Problem:

On an exam, would it be more desirable to earn a grade with a high or low percentile? Explain.

Exercise:

Problem:

Mina is waiting in line at the Department of Motor Vehicles (DMV). Her wait time of 32 minutes is the 85th percentile of wait times. Is that good or bad? Write a sentence interpreting the 85th percentile in the context of this situation.

Solution:

When waiting in line at the DMV, the 85th percentile would be a long wait time compared to the other people waiting. 85% of people had shorter wait times than Mina. In this context, Mina would prefer a wait time corresponding to a lower percentile. 85% of people at the DMV waited 32 minutes or less. 15% of people at the DMV waited 32 minutes or longer.

Exercise:

Problem:

In a survey collecting data about the salaries earned by recent college graduates, Li found that her salary was in the 78th percentile. Should Li be pleased or upset by this result? Explain.

Exercise:

Problem:

In a study collecting data about the repair costs of damage to automobiles in a certain type of crash tests, a certain model of car had $1,700 in damage and was in the 90th percentile. Should the manufacturer and the consumer be pleased or upset by this result? Explain and write a sentence that interprets the 90th percentile in the context of this problem.

Solution:

The manufacturer and the consumer would be upset. This is a large repair cost for the damages, compared to the other cars in the sample. INTERPRETATION: 90% of the crash tested cars had damage repair costs of $1700 or less; only 10% had damage repair costs of $1700 or more.

Exercise:

Problem:

The University of California has two criteria used to set admission standards for freshman to be admitted to a college in the UC system:

a. Students' GPAs and scores on standardized tests (SATs and ACTs) are entered into a formula that calculates an "admissions index" score. The admissions index score is used to set eligibility standards intended to meet the goal of admitting the top 12% of high school students in the state. In this context, what percentile does the top 12% represent?
b. Students whose GPAs are at or above the 96th percentile of all students at their high school are eligible (called eligible in the local context), even if they are not in the top 12% of all students in the state. What percentage of students from each high school are "eligible in the local context"?

Exercise:

Problem:

Suppose that you are buying a house. You and your realtor have determined that the most expensive house you can afford is the 34th percentile. The 34th percentile of housing prices is $240,000 in the town you want to move to. In this town, can you afford 34% of the houses or 66% of the houses?

Solution:

You can afford 34% of houses. 66% of the houses are too expensive for your budget. INTERPRETATION: 34% of houses cost $240,000 or less. 66% of houses cost $240,000 or more.

Use [link] to calculate the following values:

Exercise:

Problem: First quartile = _______

Exercise:

Problem: Second quartile = median = 50th percentile = _______

Solution:

4

Exercise:

Problem: Third quartile = _______

Exercise:

Problem: Interquartile range (IQR) = _____ – _____ = _____

Solution:

6 – 4 = 2

Exercise:

Problem: 10th percentile = _______

Exercise:

Problem: 70th percentile = _______

Solution:

6

Homework

Exercise:

Problem:

The median age for U.S. blacks currently is 30.9 years; for U.S. whites it is 42.3 years.

a. Based upon this information, give two reasons why the black median age could be lower than the white median age.
b. Does the lower median age for blacks necessarily mean that blacks die younger than whites? Why or why not?

c. How might it be possible for blacks and whites to die at approximately the same age, but for the median age for whites to be higher?

Exercise:

Problem:

Six hundred adult Americans were asked by telephone poll, "What do you think constitutes a middle-class income?" The results are in the following table. Also, include left endpoint, but not the right endpoint.

Salary ($)	**Relative Frequency**
< 20,000	0.02
20,000–25,000	0.09
25,000–30,000	0.19
30,000–40,000	0.26
40,000–50,000	0.18
50,000–75,000	0.17
75,000–99,999	0.02
100,000+	0.01

a. What percentage of the survey answered "not sure"?
b. What percentage think that middle-class is from $25,000 to $50,000?
c. Construct a histogram of the data.

    i. Should all bars have the same width, based on the data? Why or why not?

   - ii. How should the <20,000 and the 100,000+ intervals be handled? Why?

- d. Find the 40th and 80th percentiles
- e. Construct a bar graph of the data

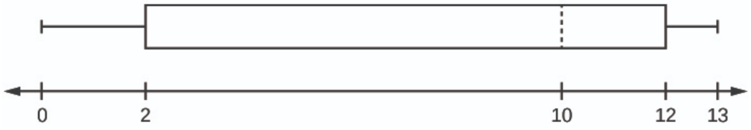
Solution:

- a. 1 – (0.02+0.09+0.19+0.26+0.18+0.17+0.02+0.01) = 0.06
- b. 0.19+0.26+0.18 = 0.63
- c. Check student’s solution.
- d. 40th percentile will fall between 30,000 and 40,000

  80th percentile will fall between 50,000 and 75,000
- e. Check student’s solution.

Exercise:

Problem: Given the following box plot:



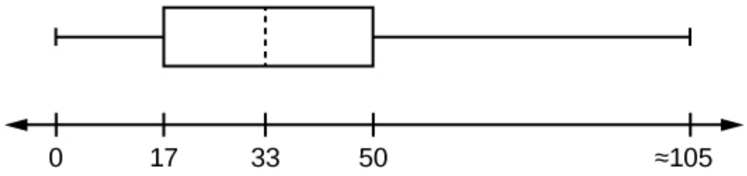
- a. Which quarter has the smallest spread of data? What is that spread?
- b. Which quarter has the largest spread of data? What is that spread?
- c. Find the interquartile range (*IQR*).
- d. Are there more data in the interval 5–10 or in the interval 10–13? How do you know this?
- e. Which interval has the fewest data in it? How do you know this?

   - i. 0–2
   - ii. 2–4
   - iii. 10–12
   - iv. 12–13
   - v. Need more information

Exercise:

Problem:

The following box plot shows the U.S. population for 1990, the latest available year.



a. Are there fewer or more children (age 17 and under) than senior citizens (age 65 and over)? How do you know?
b. 12.6% are age 65 and over. Approximately what percentage of the population are working age adults (above age 17 to age 65)?

Solution:

a. More children; the left whisker shows that 25% of the population are children 17 and younger. The right whisker shows that 25% of the population are adults 50 and older, so adults 65 and over represent less than 25%.
b. 62.4%

Glossary

Interquartile Range
: or *IQR*, is the range of the middle 50 percent of the data values; the *IQR* is found by subtracting the first quartile from the third quartile.

Outlier
: an observation that does not fit the rest of the data

Percentile
: a number that divides ordered data into hundredths; percentiles may or may not be part of the data. The median of the data is the second quartile and the 50th percentile. The first and third quartiles are the 25th and the 75th percentiles, respectively.

Quartiles

the numbers that separate the data into quarters; quartiles may or may not be part of the data. The second quartile is the median of the data.

Box Plots

Box plots (also called **box-and-whisker plots** or **box-whisker plots**) give a good graphical image of the concentration of the data. They also show how far the extreme values are from most of the data. A box plot is constructed from five values, which together make up the **five-number summary**: the minimum value, the first quartile, the median, the third quartile, and the maximum value. We use these values to compare how close other data values are to them.

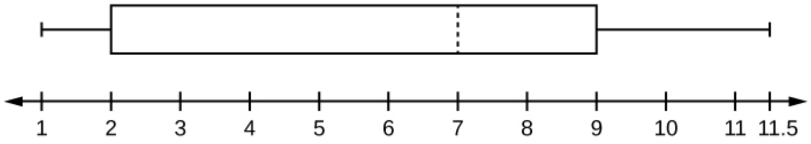
To construct a box plot, use a horizontal or vertical number line and a rectangular box. The smallest and largest data values label the endpoints of the axis. The first quartile marks one end of the box and the third quartile marks the other end of the box. Approximately **the middle 50 percent of the data fall inside the box.** The "whiskers" extend from the ends of the box to the smallest and largest data values. The median or second quartile can be between the first and third quartiles, or it can be one, or the other, or both. The box plot gives a good, quick picture of the data.

Note:
NOTE
You may encounter box-and-whisker plots that have dots marking outlier values. In those cases, the whiskers are not extending to the minimum and maximum values.

Consider, again, this dataset.

1 1 2 2 4 6 6.8 7.2 8 8.3 9 10 10 11.5

The first quartile is two, the median is seven, and the third quartile is nine. The smallest value is one, and the largest value is 11.5. The following image shows the constructed box plot.



The two whiskers extend from the first quartile to the smallest value and from the third quartile to the largest value. The median is shown with a dashed line.

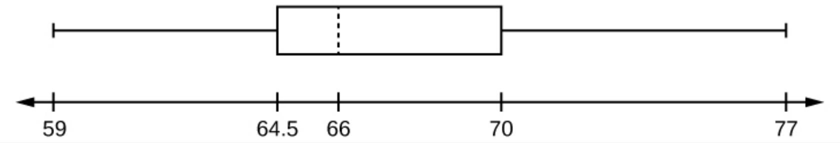
Note:
NOTE
It is important to start a box plot with a **scaled number line**. Otherwise the box plot may not be useful.

Example:
The following data are the heights of 40 students in a statistics class.

59 60 61 62 62 63 63 64 64 64 65 65 65 65 65 65 65 65 65 66 66 67 67 68 68 69 70 70 70 70 70 71 71 72 72 73 74 74 75 77

Construct a box plot of the data; calculator instructions for finding the five-number summary, as well as constructing a box plot on the calculator, follow the example.

- Minimum value = 59
- Maximum value = 77
- Q_1: First quartile = 64.5
- Q_2: Second quartile or median= 66
- Q_3: Third quartile = 70



Note the following properties from the box plot:

a. Each quarter has approximately 25% of the data.
b. The spreads of the four quarters are 64.5 – 59 = 5.5 (first quarter), 66 – 64.5 = 1.5 (second quarter), 70 – 66 = 4 (third quarter), and 77 – 70 = 7 (fourth quarter). So, the second quarter has the smallest spread and the fourth quarter has the largest spread.
c. Range = maximum value – the minimum value = 77 – 59 = 18
d. Interquartile Range: $IQR = Q_3 – Q_1 = 70 – 64.5 = 5.5$.
e. The interval 59 through 65 has more than 25% of the data so it has more data in it than the interval 66 through 70 which has 25% of the data.
f. The middle 50% (middle half) of the data has a range of 5.5 inches.

Note:

To find the five-number summary on the calculator:

Enter data into the list editor (Press STAT 1:EDIT). If you need to clear the list, arrow up to the name L1, press CLEAR, and then arrow down.

Put the data values into the list L1.

Press STAT and arrow to CALC. Press 1:1-VarStats. Enter L1.

Press ENTER.

Scroll down past n, until you can see the entire five-number summary:

minX: Smallest value = 59.

Q_1: First quartile = 64.5.

Med: Second quartile or median = 66.

Q_3: Third quartile = 70.

maxX: Largest value = 77.

To construct the box plot:

Go to STAT PLOT by hitting 2nd Y=. Make sure all other plots are turned off and then scroll to a plot number and hit ENTER. Then, hit ENTER to turn it on.

Arrow down to Type and then use the right arrow key to go to the fifth picture, which is the box plot. Press ENTER.
Arrow down to Xlist: Press 2nd 1 for L1
Arrow down to Freq: Press ALPHA 1.
Press Zoom. Press 9: ZoomStat.
Press TRACE, and use the arrow keys to examine the box plot.

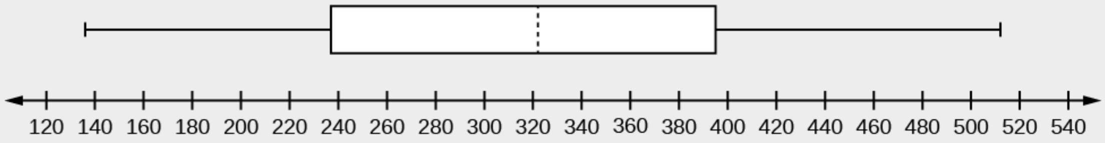
Note:
Try It
Exercise:

Problem:

The following data are the number of pages in 40 books on a shelf. Construct a box plot using a graphing calculator, and state the interquartile range.

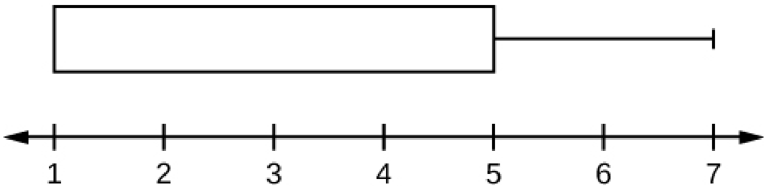
136 140 178 190 205 215 217 218 232 234 240 255 270 275 290 301 303 315 317 318 326 333 343 349 360 369 377 388 391 392 398 400 402 405 408 422 429 450 475 512

Solution:



IQR = 158

For some sets of data, some of the largest value, smallest value, first quartile, median, and third quartile may be the same. For instance, you might have a data set in which the median and the third quartile are the same. In this case, the diagram would not have a dotted line inside the box displaying the median. The right side of the box would display both the third quartile and the median. For example, if the smallest value and the first quartile were both one, the median and the third quartile were both five, and the largest value was seven, the box plot would look like:



In this case, at least 25% of the values are equal to one. Twenty-five percent of the values are between one and five, inclusive. At least 25% of the values are equal to five. The top 25% of the values fall between five and seven, inclusive.

Example:
Test scores for a college statistics class held during the day are:

99 56 78 55.5 32 90 80 81 56 59 45 77 84.5 84 70 72 68 32 79 90

Test scores for a college statistics class held during the evening are:

98 78 68 83 81 89 88 76 65 45 98 90 80 84.5 85 79 78 98 90 79 81 25.5

Exercise:

Problem:

a. Find the smallest and largest values, the median, and the first and third quartile for the day class.
b. Find the smallest and largest values, the median, and the first and third quartile for the night class.
c. For each data set, what percentage of the data is between the smallest value and the first quartile? the first quartile and the median? the median and the third quartile? the third quartile and the largest value? What percentage of the data is between the first quartile and the largest value?
d. Create a box plot for each set of data. Use one number line for both box plots.

e. Which box plot has the widest spread for the middle 50% of the data (the data between the first and third quartiles)? What does this mean for that set of data in comparison to the other set of data?

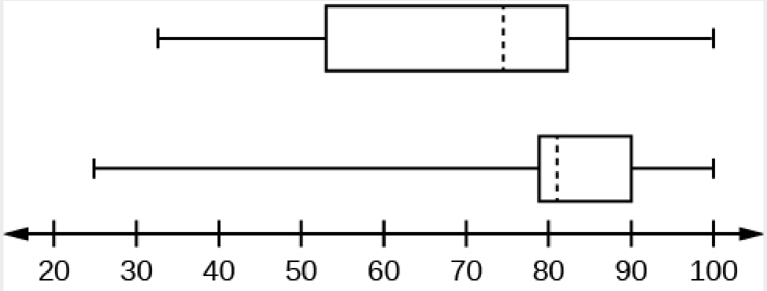
Solution:

a.
   - Min = 32
   - Q_1 = 56
   - *Med* = 74.5
   - Q_3 = 82.5
   - Max = 99

b.
   - Min = 25.5
   - Q_1 = 78
   - *Med* = 81
   - Q_3 = 89
   - Max = 98

c. Day class: There are six data values ranging from 32 to 56: 30%. There are six data values ranging from 56 to 74.5: 30%. There are five data values ranging from 74.5 to 82.5: 25%. There are five data values ranging from 82.5 to 99: 25%. There are 16 data values between the first quartile, 56, and the largest value, 99: 75%. Night class: There are six data values ranging from 25.5 and 78: 27%. There are five data values between 78 and 81: 23%. There are six data values between 81 and 89: 27%. There are five data values between 89 and 98: 23%. There are 16 data values between the first quartile, 78, and the largest value, 98: 73%.

d.



e. The first data set has the wider spread for the middle 50% of the data. The *IQR* for the first data set is greater than the *IQR* for the second set. This means that there is more variability in the middle 50% of the first data set.

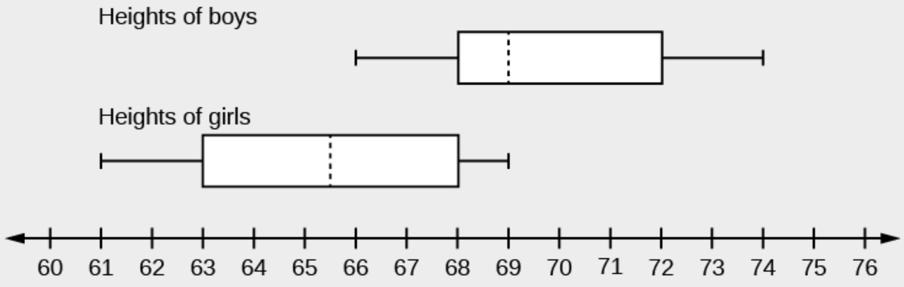
Note:
Try It
Exercise:

Problem:

The following data set shows the heights in inches for the boys in a class of 40 students.

66; 66; 67; 67; 68; 68; 68; 68; 68; 69; 69; 69; 70; 71; 72; 72; 72; 73; 73; 74
The following data set shows the heights in inches for the girls in a class of 40 students.
61; 61; 62; 62; 63; 63; 63; 65; 65; 65; 66; 66; 66; 67; 68; 68; 68; 69; 69; 69
Construct a box plot using a graphing calculator for each data set, and state which box plot has the wider spread for the middle 50% of the data.

Solution:



IQR for the boys = 4

IQR for the girls = 5

The box plot for the heights of the girls has the wider spread for the middle 50% of the data.

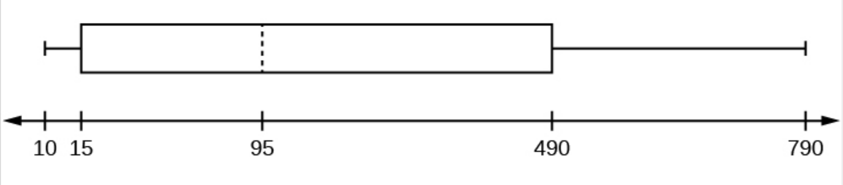
Example:
Graph a box-and-whisker plot for the data values shown.

1010101535759095100175420490515515790

The five numbers used to create a box-and-whisker plot are:

- Min: 10
- Q_1: 15
- Med: 95
- Q_3: 490
- Max: 790

The following graph shows the box-and-whisker plot.



Note:
Try It
Exercise:

Problem:

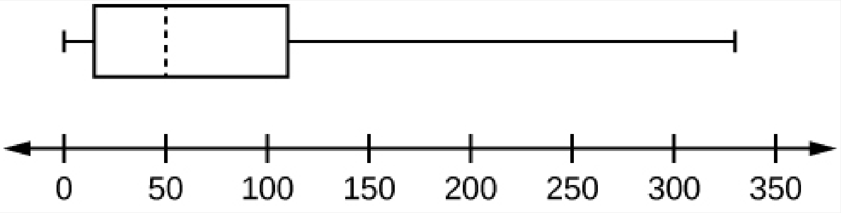
Follow the steps you used to graph a box-and-whisker plot for the data values shown.

0551530304550506075110140240330

Solution:

The data are in order from least to greatest. There are 15 values, so the eighth number in order is the median: 50. There are seven data values written to the left of the median and 7 values to the right. The five values that are used to create the boxplot are:

- Min: 0
- Q_1: 15
- Med: 50
- Q_3: 110
- Max: 330



References

Data from *West Magazine*.

Section Review

Box plots are a type of graph that can help visually organize data. To graph a box plot the following data points (or the five-number summary) must be calculated: the minimum value, the first quartile, the median, the third quartile, and the maximum value. Once the box plot is graphed, you can display and compare distributions of data.

Sixty-five randomly selected car salespersons were asked the number of cars they generally sell in one week. Fourteen people answered that they generally sell three cars; nineteen generally sell four cars; twelve generally sell five cars; nine generally sell six cars; eleven generally sell seven cars.

Exercise:

Problem:

Construct a box plot below. Use a ruler to measure and scale accurately.

Exercise:

Problem:

Looking at your box plot, does it appear that the data are concentrated together, spread out evenly, or concentrated in some areas, but not in others? How can you tell?

Solution:

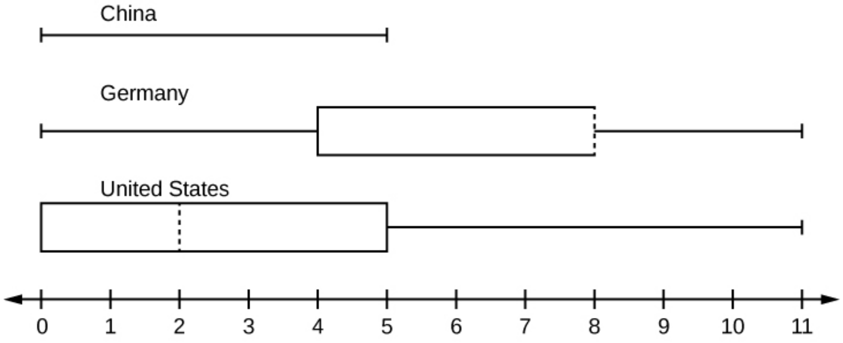
More than 25% of salespersons sell four cars in a typical week. You can see this concentration in the box plot because the first quartile is equal to the median. The top 25% and the bottom 25% are spread out evenly; the whiskers have the same length.

Homework

Exercise:

Problem:

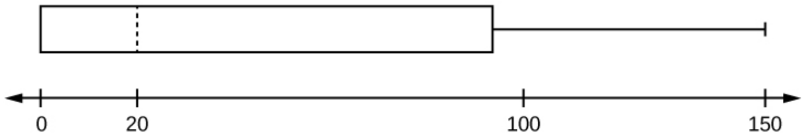
In a survey of 20-year-olds in China, Germany, and the United States, people were asked the number of foreign countries they had visited in their lifetime. The following box plots display the results.



a. In complete sentences, describe what the shape of each box plot implies about the distribution of the data collected.
b. Have more Americans or more Germans surveyed been to over eight foreign countries?
c. Compare the three box plots. What do they imply about the foreign travel of 20-year-old residents of the three countries when compared to each other?

Exercise:

Problem: Given the following box plot, answer the questions.



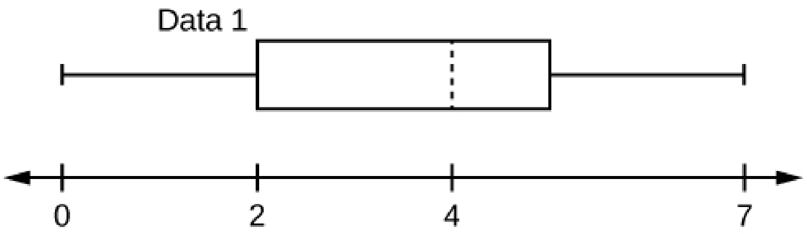
a. Think of an example (in words) where the data might fit into the above box plot. In 2–5 sentences, write down the example.
b. What does it mean to have the first and second quartiles so close together, while the second to third quartiles are far apart?

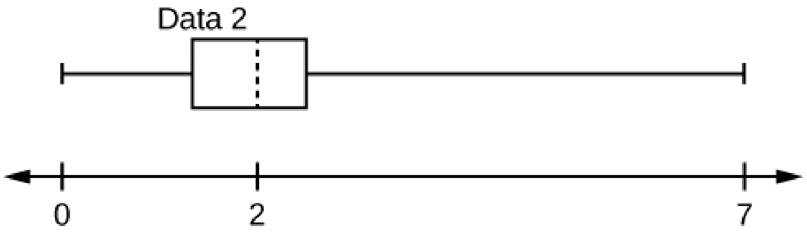
Solution:

a. Answers will vary. Possible answer: State University conducted a survey to see how involved its students are in community service. The box plot shows the number of community service hours logged by participants over the past year.
b. Because the first and second quartiles are close, the data in this quarter is very similar. There is not much variation in the values. The data in the third quarter is much more variable, or spread out. This is clear because the second quartile is so far away from the third quartile.

Exercise:

Problem: Given the following box plots, answer the questions.





a. In complete sentences, explain why each statement is false.

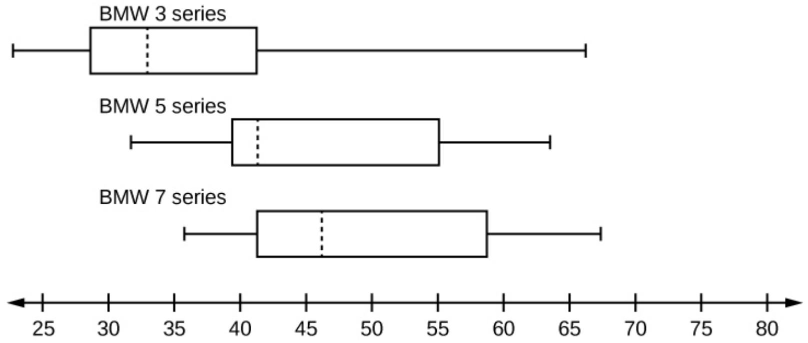
   i. **Data 1** has more data values above two than **Data 2** has above two.
   ii. The data sets cannot have the same mode.
   iii. For **Data 1**, there are more data values below four than there are above four.

b. For which group, Data 1 or Data 2, is the value of “7” more likely to be an outlier? Explain why in complete sentences.

Exercise:

Problem:

A survey was conducted of 130 purchasers of new BMW 3 series cars, 130 purchasers of new BMW 5 series cars, and 130 purchasers of new BMW 7 series cars. In it, people were asked the age they were when they purchased their car. The following box plots display the results.



a. In complete sentences, describe what the shape of each box plot implies about the distribution of the data collected for that car series.
b. Which group is most likely to have an outlier? Explain how you determined that.
c. Compare the three box plots. What do they imply about the age of purchasing a BMW from the series when compared to each other?
d. Look at the BMW 5 series. Which quarter has the smallest spread of data? What is the spread?
e. Look at the BMW 5 series. Which quarter has the largest spread of data? What is the spread?
f. Look at the BMW 5 series. Estimate the interquartile range (IQR).
g. Look at the BMW 5 series. Are there more data in the interval 31 to 38 or in the interval 45 to 55? How do you know this?
h. Look at the BMW 5 series. Which interval has the fewest data in it? How do you know this?

    i. 31–35
    ii. 38–41
    iii. 41–64

Solution:

a. Each box plot is spread out more in the greater values. Each plot is skewed to the right, so the ages of the top 50% of buyers are more variable than the ages of the lower 50%.

b. The BMW 3 series is most likely to have an outlier. It has the longest whisker.
c. Comparing the median ages, younger people tend to buy the BMW 3 series, while older people tend to buy the BMW 7 series. However, this is not a rule, because there is so much variability in each data set.
d. The second quarter has the smallest spread. There seems to be only a three-year difference between the first quartile and the median.
e. The third quarter has the largest spread. There seems to be approximately a 14-year difference between the median and the third quartile.
f. *IQR* ~ 17 years
g. There is not enough information to tell. Each interval lies within a quarter, so we cannot tell exactly where the data in that quarter is concentrated.
h. The interval from 31 to 35 years has the fewest data values. Twenty-five percent of the values fall in the interval 38 to 41, and 25% fall between 41 and 64. Since 25% of values fall between 31 and 38, we know that fewer than 25% fall between 31 and 35.

Exercise:

Problem:

Twenty-five randomly selected students were asked the number of movies they watched the previous week. The results are as follows:

# of movies	Frequency
0	5

# of movies	Frequency
1	9
2	6
3	4
4	1

Construct a box plot of the data.

Bringing It Together

Exercise:

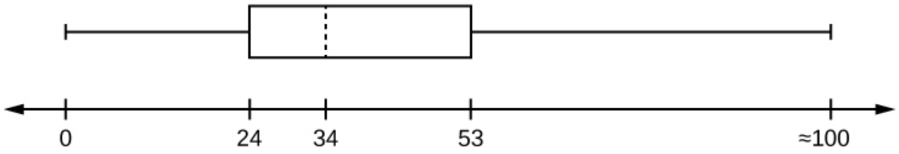
Problem:

Santa Clara County, CA, has approximately 27,873 Japanese-Americans. Their ages are as follows:

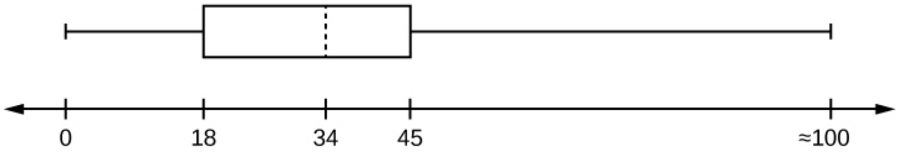
Age Group	Percent of Community
0–17	18.9
18–24	8.0
25–34	22.8
35–44	15.0

Age Group	Percent of Community
45–54	13.1
55–64	11.9
65+	10.3

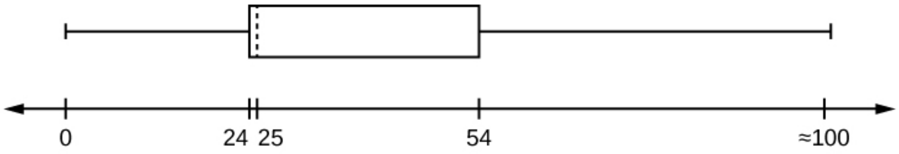
a. Construct a histogram of the Japanese-American community in Santa Clara County, CA. The bars will **not** be the same width for this example. Why not? What impact does this have on the reliability of the graph?
b. What percentage of the community is under age 35?
c. Which box plot most resembles the information above?



(a)



(b)



(c)

Solution:

a. For graph, check student's solution.
b. 49.7% of the community is under the age of 35.
c. Based on the information in the table, graph (a) most closely represents the data.

Glossary

Box plot

a graph that gives a quick picture of the middle 50% of the data

First Quartile

the value that is the median of the lower half of the ordered data set

Third Quartile

the value that is the median of the upper half of the ordered data set

Five-Number Summary

five particular numbers that summarize a data set: the minimum data value, the first quartile, the median, the third quartile, and the maximum data value.

Frequency Polygon

looks like a line graph but uses intervals to display ranges of large amounts of data

Interval

also called a class interval; an interval represents a range of data and is used when displaying large data sets

Measures of the Center of the Data

The "center" of a data set is also a way of describing location. The two most widely used measures of the "center" of the data are the **mean** (average) and the **median**. To calculate the **mean weight** of 50 people, add the 50 weights together and divide by 50. To find the **median weight** of the 50 people, order the data and find the number that splits the data into two equal parts. The median is generally a better measure of the center when there are extreme values or outliers because it is not affected by the precise numerical values of the outliers. The mean is the most common measure of the center.

Note:
NOTE
The words “mean” and “average” are often used interchangeably. The substitution of one word for the other is common practice. The technical term is “arithmetic mean” and “average” is technically a center location. However, in practice among non-statisticians, “average" is commonly accepted for “arithmetic mean.”

When each value in the data set is not unique, the mean can be calculated by multiplying each distinct value by its frequency and then dividing the sum by the total number of data values. The letter used to represent the **sample mean** is an x with a bar over it (pronounced “x bar”): x.

The Greek letter μ (pronounced "mew") represents the **population mean**. One of the requirements for the **sample mean** to be a good estimate of the **population mean** is for the sample taken to be truly random.

To see that both ways of calculating the mean are the same, consider the sample:
1; 1; 1; 2; 2; 3; 4; 4; 4; 4; 4
Equation:

$$x = \frac{1+1+1+2+2+3+4+4+4+4+4}{11} = 2.7$$

Equation:

$$x = \frac{3(1)+2(2)+1(3)+5(4)}{11} = 2.7$$

In the second calculation, the frequencies are 3, 2, 1, and 5.

You can quickly find the location of the median by using the expression $\frac{n+1}{2}$.

The letter n is the total number of data values in the sample. If n is an odd number, the median is the middle value of the ordered data (ordered smallest to largest). If n is an even number, the median is equal to the two middle values added together and divided by two after the data has been ordered. For example, if the total number of data values is 97, then $\frac{n+1}{2} = \frac{97+1}{2} = 49$. The median is the 49th value in the ordered data. If the total number of data values is 100, then $\frac{n+1}{2} = \frac{100+1}{2} = 50.5$. The median occurs midway between the 50th and 51st values. The location of the median and the value of the median are **not** the same. The upper case letter M is often used to represent the median. The next example illustrates the location of the median and the value of the median.

Example:
Exercise:

Problem:

AIDS data indicating the number of months a patient with AIDS lives after taking a new antibody drug are as follows (smallest to largest):

3; 4; 8; 8; 10; 11; 12; 13; 14; 15; 15; 16; 16; 17; 17; 18; 21; 22; 22; 24; 24; 25; 26; 26; 27; 27; 29; 29; 31; 32; 33; 33; 34; 34; 35; 37; 40; 44; 44; 47;

Calculate the mean and the median.

Solution:

The calculation for the mean is:

$$x = \frac{[3+4+(8)(2)+10+11+12+13+14+(15)(2)+(16)(2)+...+35+37+40+(44)(2)+47]}{40} = 23.6$$

To find the median, M, first use the formula for the location. The location is:

$$\frac{n+1}{2} = \frac{40+1}{2} = 20.5$$

Starting at the smallest value, the median is located between the 20th and 21st values (the two 24s):

3; 4; 8; 8; 10; 11; 12; 13; 14; 15; 15; 16; 16; 17; 17; 18; 21; 22; 22; 24; 24; 25; 26; 26; 27; 27; 29; 29; 31; 32; 33; 33; 34; 34; 35; 37; 40; 44; 44; 47;

$$M = \frac{24+24}{2} = 24$$

Note:
To find the mean and the median:
Clear list L1. Pres STAT 4:ClrList. Enter 2nd 1 for list L1. Press ENTER.
Enter data into the list editor. Press STAT 1:EDIT.
Put the data values into list L1.
Press STAT and arrow to CALC. Press 1:1-VarStats. Press 2nd 1 for L1 and then ENTER.
Press the down and up arrow keys to scroll.
x = 23.6, M = 24

Note:
Try It
Exercise:

Problem:

The following data show the number of months patients typically wait on a transplant list before getting surgery. The data are ordered from smallest to largest. Calculate the mean and median.

3 4 5 7 7 7 7 8 8 9 9 10 10 10 10 10 11 12 12 13 14 14 15 15 17 17 18 19 19 19 21 21 22 22 23 24 24 24 24

Solution:

Mean: 3 + 4 + 5 + 7 + 7 + 7 + 7 + 8 + 8 + 9 + 9 + 10 + 10 + 10 + 10 + 10 + 11 + 12 + 12 + 13 + 14 + 14 + 15 + 15 + 17 + 17 + 18 + 19 + 19 + 19 + 21 + 21 + 22 + 22 + 23 + 24 + 24 + 24 = 544

$\frac{544}{39} = 13.95$

Median: Starting at the smallest value, the median is the 20th term, which is 13.

Example:
Exercise:

Problem:

Suppose that in a small town of 50 people, one person earns $5,000,000 per year and the other 49 each earn $30,000. Which is the better measure of the "center": the mean or the median?

Solution:

$x = \frac{5{,}000{,}000+49(30{,}000)}{50} = 129{,}400$

M = 30,000

(There are 49 people who earn $30,000 and one person who earns $5,000,000.)

The median is a better measure of the "center" than the mean because 49 of the values are 30,000 and one is 5,000,000. The 5,000,000 is an outlier. The 30,000 gives us a better sense of the middle of the data.

Note:
Try It
Exercise:

Problem:

In a sample of 60 households, one house is worth $2,500,000. Half of the rest are worth $280,000, and all the others are worth $315,000. Which is the better measure of the “center”: the mean or the median?

Solution:

The median is the better measure of the “center” than the mean because 59 of the values are $280,000 and one is $2,500,000. The $2,500,000 is an outlier. Either $280,000 or $315,000 gives us a better sense of the middle of the data.

Another measure of the center is the mode. The **mode** is the most frequent value. There can be more than one mode in a data set as long as those values have the same frequency and that frequency is the highest. A data set with two modes is called bimodal.

Example:
Statistics exam scores for 20 students are as follows:
50535959636372727272727678818384849093
Exercise:

Problem: Find the mode.

Solution:

The most frequent score is 72, which occurs five times. Mode = 72.

Note:
Try It
Exercise:

Problem: The number of books checked out from the library from 25 students are as follows:

0001233445577778889101011111212
Find the mode.

Solution:

The most frequent number of books is 7, which occurs four times. Mode = 7.

Example:

Five real estate exam scores are 430, 430, 480, 480, 495. The data set is bimodal because the scores 430 and 480 each occur twice.

When is the mode the best measure of the "center"? Consider a weight loss program that advertises a mean weight loss of six pounds the first week of the program. The mode might indicate that most people lose two pounds the first week, making the program less appealing.

Note:
NOTE
The mode can be calculated for qualitative data as well as for quantitative data. For example, if the data set is: red, red, red, green, green, yellow, purple, black, blue, the mode is red.

Statistical software will easily calculate the mean, the median, and the mode. Some graphing calculators can also make these calculations. In the real world, people make these calculations using software.

Note:
Try It
Exercise:

Problem:

Five credit scores are 680, 680, 700, 720, 720. The data set is bimodal because the scores 680 and 720 each occur twice. Consider the annual earnings of workers at a factory. The mode is $25,000 and occurs 150 times out of 301. The median is $50,000 and the mean is $47,500. What would be the best measure of the “center”?

Solution:

Because $25,000 occurs nearly half the time, the mode would be the best measure of the center because the median and mean don’t represent what most people make at the factory.

The Law of Large Numbers and the Mean

The Law of Large Numbers says that if you take samples of larger and larger size from any population, then the mean x of the sample is very likely to get closer and closer to μ. This is discussed in more detail later in the text.

Sampling Distributions and Statistic of a Sampling Distribution

You can think of a **sampling distribution** as a **relative frequency distribution** with a great many samples. (See **Sampling and Data** for a review of relative frequency). Suppose thirty randomly selected students were asked the number of movies they watched the previous week. The results are in the **relative frequency table** shown below.

# of movies	**Relative Frequency**
0	$\frac{5}{30}$
1	$\frac{15}{30}$
2	$\frac{6}{30}$
3	$\frac{3}{30}$
4	$\frac{1}{30}$

If you let the number of samples get very large (say, 300 million or more), the relative frequency table becomes a relative frequency distribution.

A **statistic** is a number calculated from a sample. Statistic examples include the mean, the median and the mode as well as others. The sample mean x is an example of a statistic which estimates the population mean μ.

Calculating the Mean of Grouped Frequency Tables

When only grouped data is available, you do not know the individual data values (we only know intervals and interval frequencies); therefore, you cannot compute an exact mean for the data set. What we must do is estimate the actual mean by calculating the mean of a frequency table. A frequency table is a data representation in which grouped data is displayed along with the corresponding frequencies. To calculate the mean from a grouped frequency table we can apply the basic definition of mean: *mean* = $\frac{data\ sum}{number\ of\ data\ values}$ We simply need to modify the definition to fit within the restrictions of a frequency table.

Since we do not know the individual data values we can instead find the midpoint of each interval. The midpoint is $\frac{lower\ boundary+upper\ boundary}{2}$. We can now modify the mean definition to be $Mean\ of\ Frequency\ Table = \frac{\sum fm}{\sum f}$ where f = the frequency of the interval and m = the midpoint of the interval.

Example:
Exercise:

Problem:

A frequency table displaying professor Blount’s last statistic test is shown. Find the best estimate of the class mean.

Grade Interval	**Number of Students**
50–56.5	1
56.5–62.5	0
62.5–68.5	4
68.5–74.5	4
74.5–80.5	2
80.5–86.5	3
86.5–92.5	4
92.5–98.5	1

Solution:

- Find the midpoints for all intervals

Grade Interval	**Midpoint**
50–56.5	53.25
56.5–62.5	59.5
62.5–68.5	65.5
68.5–74.5	71.5
74.5–80.5	77.5

Grade Interval	**Midpoint**
80.5–86.5	83.5
86.5–92.5	89.5
92.5–98.5	95.5

- Calculate the sum of the product of each interval frequency and midpoint. $\sum fm$

  $53.25(1) + 59.5(0) + 65.5(4) + 71.5(4) + 77.5(2) + 83.5(3) + 89.5(4) + 95.5(1) = 1460.25$
- $\mu = \frac{\sum fm}{\sum f} = \frac{1460.25}{19} = 76.86$

Note:
Try It
Exercise:

Problem:

Maris conducted a study on the effect that playing video games has on memory recall. As part of her study, she compiled the following data:

Hours Teenagers Spend on Video Games	**Number of Teenagers**
0–3.5	3
3.5–7.5	7
7.5–11.5	12
11.5–15.5	7
15.5–19.5	9

What is the best estimate for the mean number of hours spent playing video games?

Solution:

Find the midpoint of each interval, multiply by the corresponding number of teenagers, add the results and then divide by the total number of teenagers

The midpoints are 1.75, 5.5, 9.5, 13.5, 17.5.
Mean = ((1.75)(3) + (5.5)(7) + (9.5)(12) + (13.5)(7) + (17.5)(9))/38 = 409.75/38 = 10.78

References

Data from The World Bank, available online at http://www.worldbank.org (accessed April 3, 2013).

“Demographics: Obesity – adult prevalence rate.” Indexmundi. Available online at http://www.indexmundi.com/g/r.aspx?t=50&v=2228&l=en (accessed April 3, 2013).

Section Review

The mean and the median can be calculated to help you find the "center" of a data set. The mean is the best estimate for the actual data set, but the median is the best measurement when a data set contains several outliers or extreme values. The mode will tell you the most frequently occuring datum (or data) in your data set. The mean, median, and mode are extremely helpful when you need to analyze your data, but if your data set consists of ranges which lack specific values, the mean may seem impossible to calculate. However, the mean can be approximated if you add the lower boundary with the upper boundary and divide by two to find the midpoint of each interval. Multiply each midpoint by the number of values found in the corresponding range. Divide the sum of these values by the total number of data values in the set.

Formula Review

$\mu = \frac{\sum fm}{\sum f}$ Where f = interval frequencies and m = interval midpoints.

Exercise:

Problem: Find the mean for the following frequency tables.

a.

Grade	**Frequency**
49.5–59.5	2
59.5–69.5	3
69.5–79.5	8
79.5–89.5	12
89.5–99.5	5

b.

Daily Low Temperature	**Frequency**
49.5–59.5	53
59.5–69.5	32
69.5–79.5	15
79.5–89.5	1
89.5–99.5	0

c.

Points per Game	**Frequency**
49.5–59.5	14
59.5–69.5	32
69.5–79.5	15
79.5–89.5	23
89.5–99.5	2

Use the following information to answer the next three exercises: The following data show the lengths of boats moored in a marina. The data are ordered from smallest to largest:
161719202021232425252526262727272829303233333435373940

Exercise:

Problem: Calculate the mean.

Solution:

Mean: 16 + 17 + 19 + 20 + 20 + 21 + 23 + 24 + 25 + 25 + 25 + 26 + 26 + 27 + 27 + 27 + 28 + 29 + 30 + 32 + 33 + 33 + 34 + 35 + 37 + 39 + 40 = 738;

$\frac{738}{27}$ = 27.33

Exercise:

Problem: Identify the median.

Exercise:

Problem: Identify the mode.

Solution:

The most frequent lengths are 25 and 27, which occur three times. Mode = 25, 27

Use the following information to answer the next three exercises: Sixty-five randomly selected car salespersons were asked the number of cars they generally sell in one week. Fourteen people answered that they generally sell three cars; nineteen generally sell four cars; twelve generally sell five cars; nine generally sell six cars; eleven generally sell seven cars. Calculate the following:

Exercise:

Problem: sample mean = x = _______

Exercise:

Problem: median = _______

Solution:

4

Exercise:

Problem: mode = _______

Homework

Exercise:

Problem:

The most obese countries in the world have obesity rates that range from 11.4% to 74.6%. This data is summarized in the following table.

Percent of Population Obese	Number of Countries
11.4–20.45	29
20.45–29.45	13
29.45–38.45	4
38.45–47.45	0

Percent of Population Obese	**Number of Countries**
47.45–56.45	2
56.45–65.45	1
65.45–74.45	0
74.45–83.45	1

a. What is the best estimate of the average obesity percentage for these countries?
b. The United States has an average obesity rate of 33.9%. Is this rate above average or below?
c. How does the United States compare to other countries?

Exercise:

Problem:

The following table gives the percent of children under five considered to be underweight. What is the best estimate for the mean percentage of underweight children?

Percent of Underweight Children	**Number of Countries**
16–21.45	23
21.45–26.9	4
26.9–32.35	9
32.35–37.8	7
37.8–43.25	6
43.25–48.7	1

Solution:

The mean percentage, $x = \frac{1328.65}{50} = 26.57$

Bringing It Together

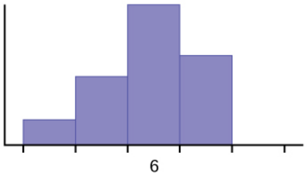
Exercise:

Problem:

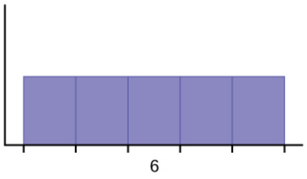
Javier and Ercilia are supervisors at a shopping mall. Each was given the task of estimating the mean distance that shoppers live from the mall. They each randomly surveyed 100 shoppers. The samples yielded the following information.

	Javier	**Ercilia**
x	6.0 miles	6.0 miles
s	4.0 miles	7.0 miles

a. How can you determine which survey was correct ?
b. Explain what the difference in the results of the surveys implies about the data.
c. If the two histograms depict the distribution of values for each supervisor, which one depicts Ercilia's sample? How do you know?

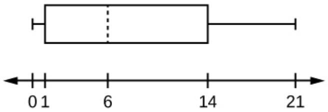


(a)

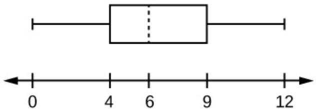


(b)

d. If the two box plots depict the distribution of values for each supervisor, which one depicts Ercilia’s sample? How do you know?



(a)



(b)

Use the following information to answer the next three exercises: We are interested in the number of years students in a particular elementary statistics class have lived in California. The information in the following table is from the entire section.

Number of years	**Frequency**	**Number of years**	**Frequency**
7	1	22	1

			Total = 20

Number of years	**Frequency**	**Number of years**	**Frequency**
14	3	23	1
15	1	26	1
18	1	40	2
19	4	42	2
20	3		
			Total = 20

Exercise:

Problem: What is the *IQR*?

a. 8
b. 11
c. 15
d. 35

Solution:

a

Exercise:

Problem: What is the mode?

a. 19
b. 19.5
c. 14 and 20
d. 22.65

Exercise:

Problem: Is this a sample or the entire population?

a. sample
b. entire population
c. neither

Solution:

b

Glossary

Frequency Table

a data representation in which grouped data is displayed along with the corresponding frequencies

Mean

a number that measures the central tendency of the data; a common name for mean is 'average.' The term 'mean' is a shortened form of 'arithmetic mean.' By definition, the mean for a sample (denoted by x) is $x \ = \ \frac{\text{Sum of all values in the sample}}{\text{Number of values in the sample}}$, and the mean for a population (denoted by μ) is $\mu = \frac{\text{Sum of all values in the population}}{\text{Number of values in the population}}$.

Median

a number that separates ordered data into halves; half the values are the same number or smaller than the median and half the values are the same number or larger than the median. The median may or may not be part of the data.

Midpoint

the mean of an interval in a frequency table

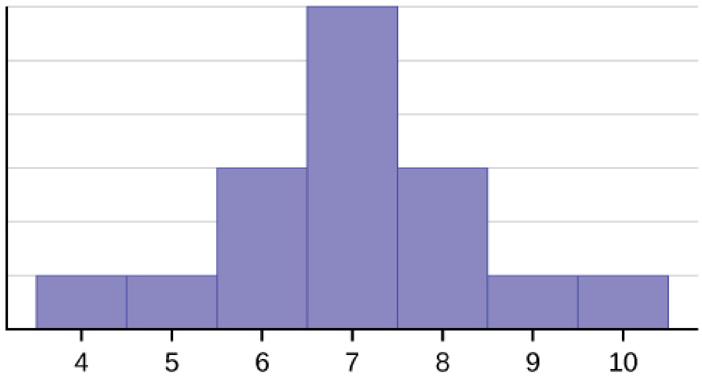
Mode

the value that appears most frequently in a set of data

Skewness and the Mean, Median, and Mode

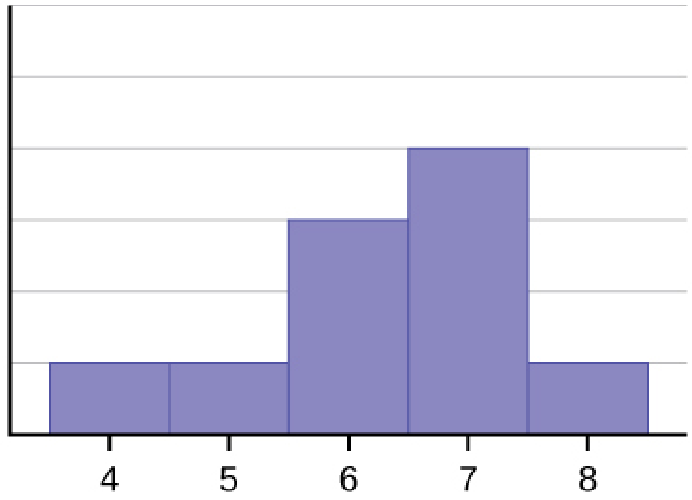
Consider the following data set.
4; 5; 6; 6; 6; 7; 7; 7; 7; 7; 7; 8; 8; 8; 9; 10

This data set can be represented by following histogram. Each interval has width one, and each value is located in the middle of an interval.



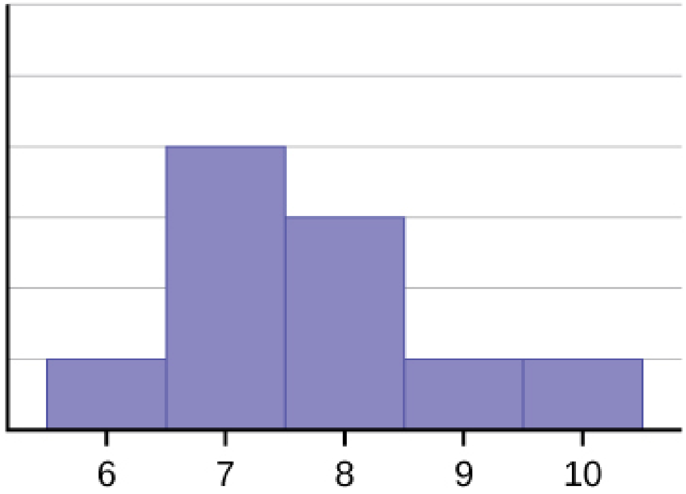
The histogram displays a **symmetrical** distribution of data. A distribution is symmetrical if a vertical line can be drawn at some point in the histogram such that the shape to the left and the right of the vertical line are mirror images of each other. The mean, the median, and the mode are each seven for these data. **In a perfectly symmetrical distribution, the mean and the median are the same.** This example has one mode (unimodal), and the mode is the same as the mean and median. In a symmetrical distribution that has two modes (bimodal), the two modes would be different from the mean and median.

The histogram for the data: 4566677778 is not symmetrical. The right-hand side seems "chopped off" compared to the left side. A distribution of this type is called **skewed to the left** because it is more stretched out to the left. That is, the lower values are more spread out.



The mean is 6.3, the median is 6.5, and the mode is seven. **Notice that the mean is less than the median, and they are both less than the mode.** The mean and the median both reflect the skewing, but the mean reflects it more so.

The histogram for the data: 67777888910, is also not symmetrical. Notice how it's more stretched out to the right. It is **skewed to the right**.

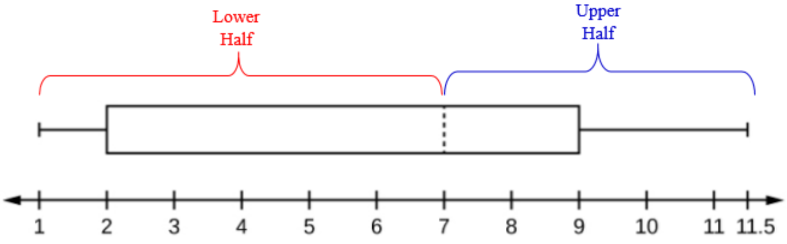


The mean is 7.7, the median is 7.5, and the mode is seven. Of the three statistics, **the mean is the largest, while the mode is the smallest**. Again, the mean reflects the skewing the most.

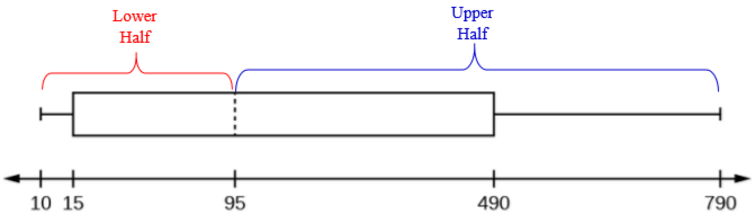
To summarize, generally if the distribution of data is skewed to the left, the mean is less than the median, which is often less than the mode. If the distribution of data is skewed to the right, the mode is often less than the median, which is less than the mean.

Skewness can also be easily determined by looking at a box plot. To do this, we compare the ranges of the lower and upper halves of the data. The range of the lower half is the distance between the minimum data value and the median. The range of the upper half is the distance between the maximum data value and the median.

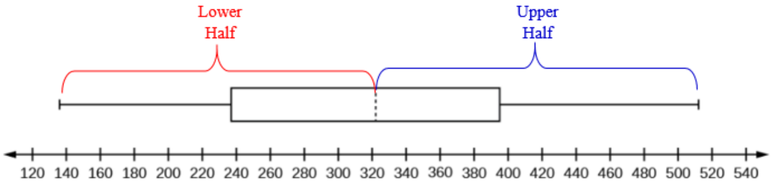
If the range of the lower half is noticeably larger than the range of the upper half, then the data is **skewed to the left**. If the range of the upper half is noticeably larger than the range of the lower half, then the data is **skewed to the right**. If the ranges are roughly the same, then the data is fairly **symmetric**.



This data is skewed to the left.



This data is skewed to the right.



This data is symmetric.

Skewness and symmetry become important when we discuss probability distributions in later chapters.

Example:
Exercise:

Problem:

Statistics are used to compare and sometimes identify authors. The following lists shows a simple random sample that compares the letter counts for three authors.

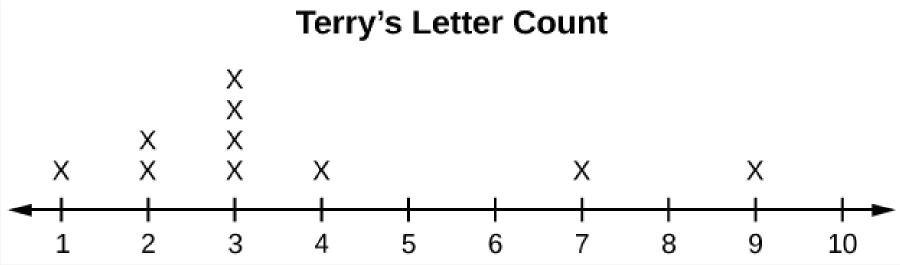
Terry: 7; 9; 3; 3; 3; 4; 1; 3; 2; 2

Davis: 3; 3; 3; 4; 1; 4; 3; 2; 3; 1

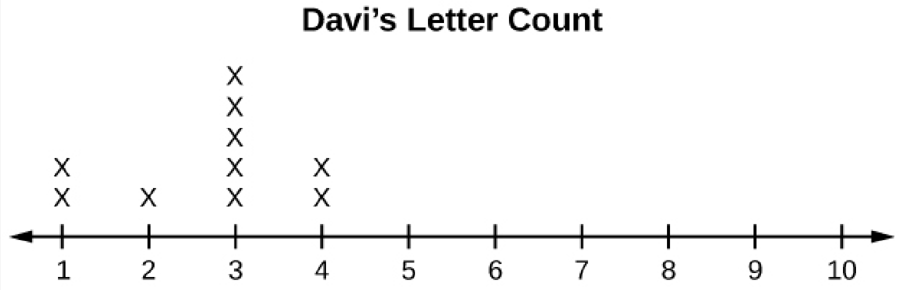
Maris: 2; 3; 4; 4; 4; 6; 6; 6; 8; 3

a. Make a dot plot for the three authors and compare the shapes.
b. Calculate the mean for each.
c. Calculate the median for each.
d. Describe any pattern you notice between the shape and the measures of center.

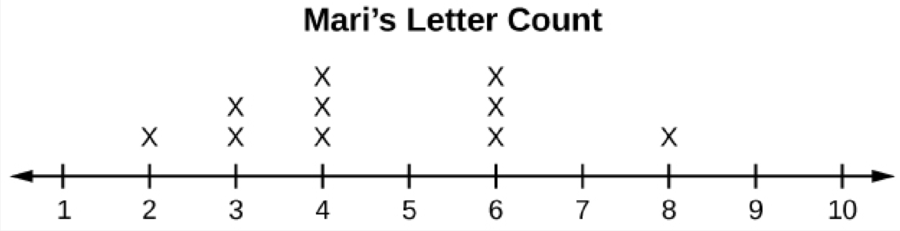
Solution:

a. 

Terry’s distribution has a right (positive) skew.



Davis’ distribution has a left (negative) skew



Maris’ distribution is symmetrically shaped.

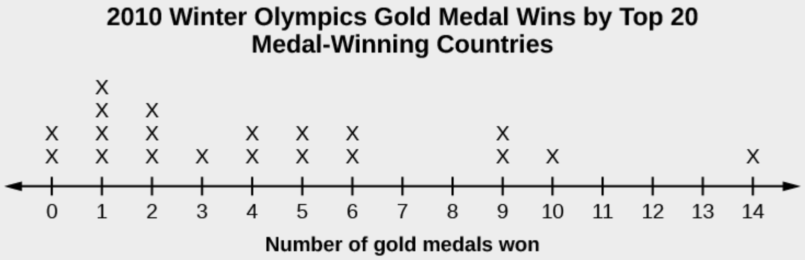
b. Terry’s mean is 3.7, Davis’ mean is 2.7, Maris’ mean is 4.6.
c. Terry’s median is three, Davis’ median is three. Maris’ median is four.
d. It appears that the median is always closest to the high point (the mode), while the mean tends to be farther out on the tail. In a symmetrical distribution, the mean and the median are both centrally located close to the high point of the distribution.

Note:
Try It
Exercise:

Problem:

Discuss the mean, median, and mode for each of the following problems. Is there a pattern between the shape and measure of the center?

a.

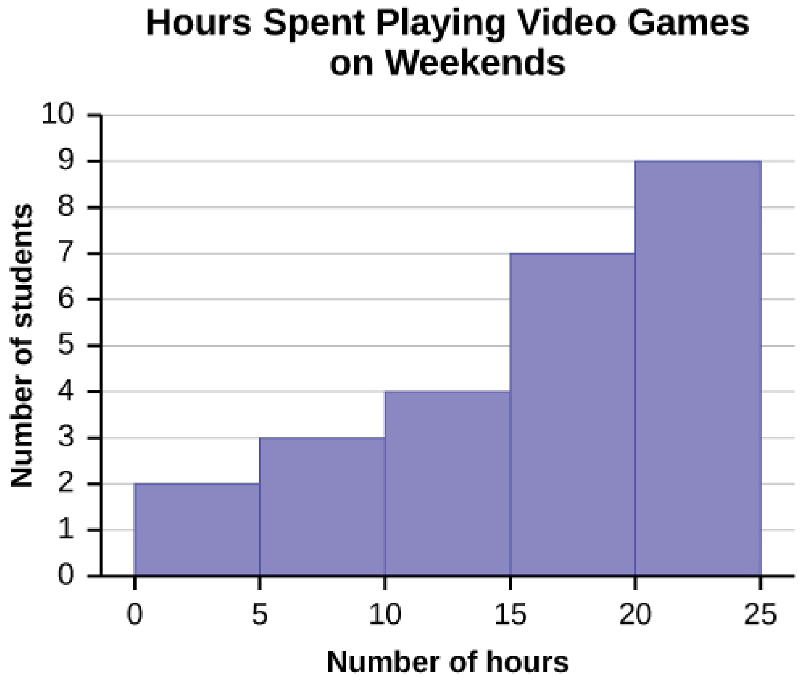


b.

The Ages Former U.S Presidents Died	
4	6 9
5	3 6 7 7 7 8
6	0 0 3 3 4 4 5 6 7 7 7 8
7	0 1 1 2 3 4 7 8 8 9

The Ages Former U.S Presidents Died	
8	0 1 3 5 8
9	0 0 3 3
Key: 8|0 means 80.	

c.



Solution:

a. mean = 4.25, median = 3.5, mode = 1; The mean > median > mode which indicates skewness to the right. (data are 0, 1, 2, 3, 4, 5, 6, 9, 10, 14 and respective frequencies are 2, 4, 3, 1, 2, 2, 2, 2, 1, 1)
b. mean = 70.1 , median = 68, mode = 57, 67 bimodal; the mean and median are close but there is a little skewness to the right which is influenced by the data being bimodal. (data are 46, 49, 53, 56, 57, 57, 57, 58, 60, 60, 63, 63, 64, 64, 65, 66, 67, 67, 67,

68, 70, 71, 71, 72, 73, 74, 77, 78, 78, 79, 80, 81, 83, 85, 88, 90, 90 93, 93).

c. These are estimates: mean = 16.1, median = 17.5, mode = 22.5 (it's possible that there is no mode); The mean < median < mode, which indicates skewness to the left. (Data used to make estimates are the midpoints of the intervals: 2.5, 7.5, 12.5, 17.5, 22.5 and respective frequencies are 2, 3, 4, 7, 9).

Section Review

Looking at the distribution of data can reveal a lot about the relationship between the mean, the median, and the mode. There are three types of distributions. A **right (or positive) skewed** distribution has a shape like [link]. A **left (or negative) skewed** distribution has a shape like [link]. A **symmetrical** distribution looks like [link].

Use the following information to answer the next three exercises: State whether the data are symmetrical, skewed to the left, or skewed to the right.

Exercise:

Problem: 1112222333333344455

Solution:

The data are symmetrical. The median is 3 and the mean is 2.85. They are close, and the mode lies close to the middle of the data, so the data are symmetrical.

Exercise:

Problem:16171922222222223

Exercise:

Problem:8787878787878889899091

Solution:

The data are skewed right. The median is 87.5 and the mean is 88.2. Even though they are close, the mode lies to the left of the middle of the data, and there are many more instances of 87 than any other number, so the data are skewed right.

Exercise:

Problem:

When the data are skewed left, what is the typical relationship between the mean and median?

Exercise:

Problem:

When the data are symmetrical, what is the typical relationship between the mean and median?

Solution:

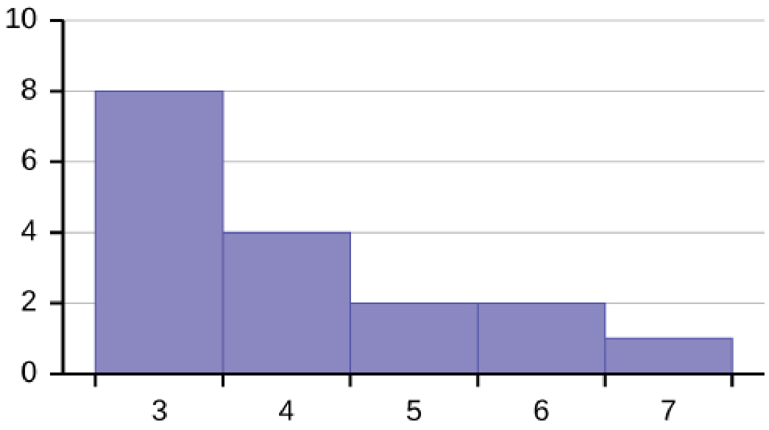
When the data are symmetrical, the mean and median are close or the same.

Exercise:

Problem: What word describes a distribution that has two modes?

Exercise:

Problem: Describe the shape of this distribution.



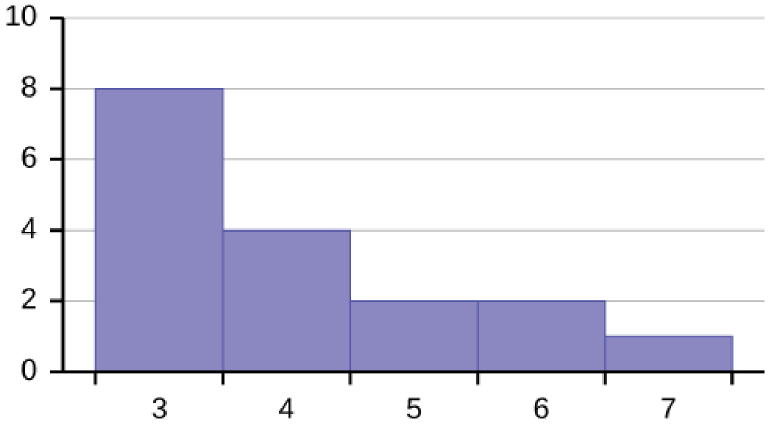
Solution:

The distribution is skewed right because it looks pulled out to the right.

Exercise:

Problem:

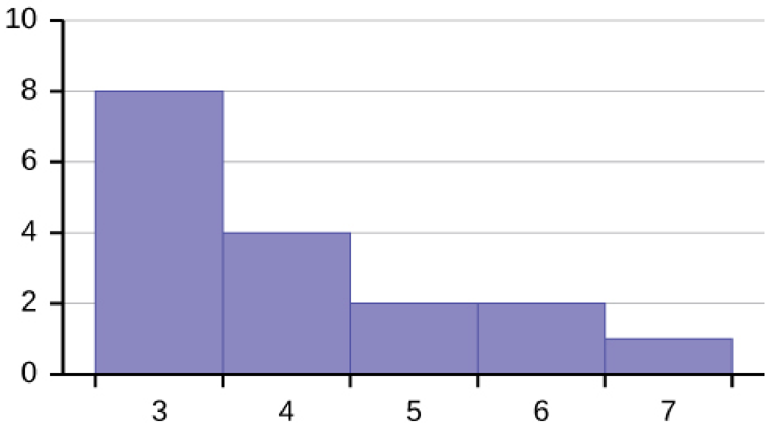
Describe the relationship between the mode and the median of this distribution.



Exercise:

Problem:

Describe the relationship between the mean and the median of this distribution.

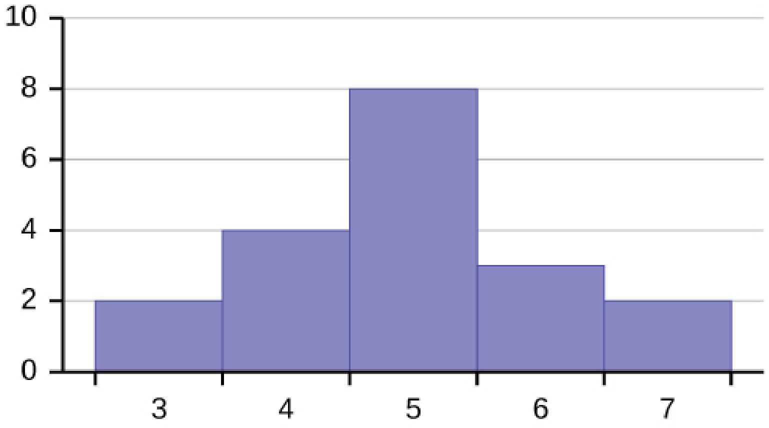


Solution:

The mean is 4.1 and is slightly greater than the median, which is four.

Exercise:

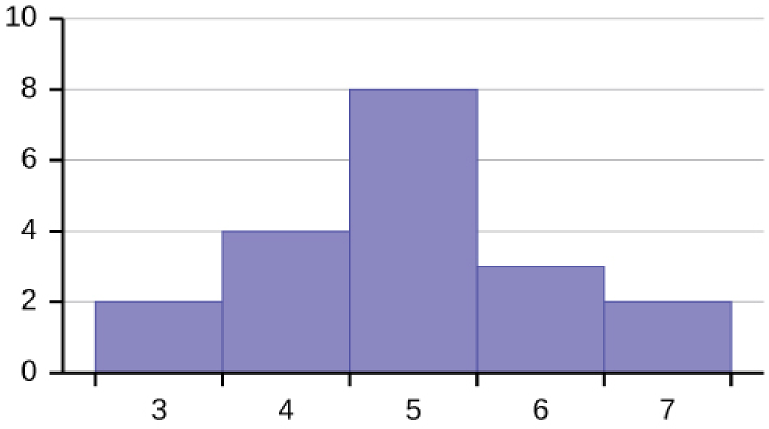
Problem: Describe the shape of this distribution.



Exercise:

Problem:

Describe the relationship between the mode and the median of this distribution.



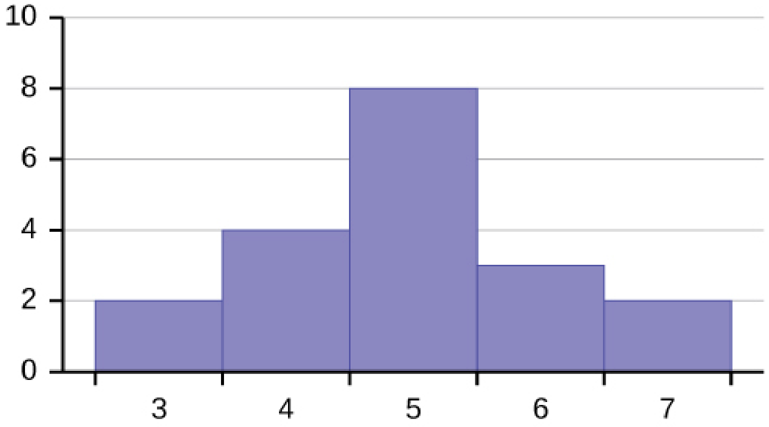
Solution:

The mode and the median are the same. In this case, they are both five.

Exercise:

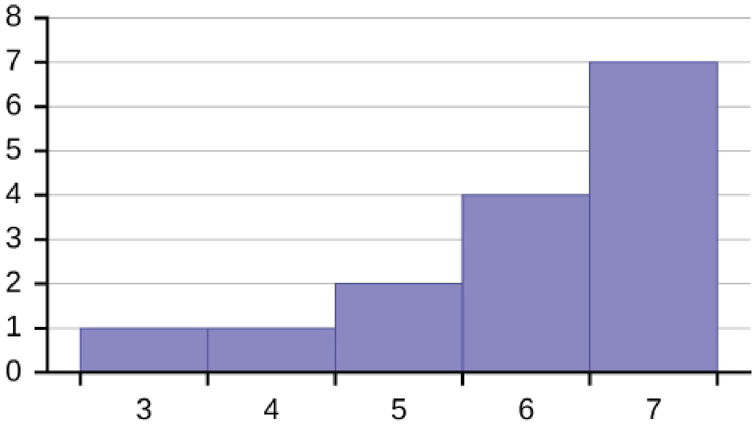
Problem:

Are the mean and the median the exact same in this distribution? Why or why not?



Exercise:

Problem: Describe the shape of this distribution.



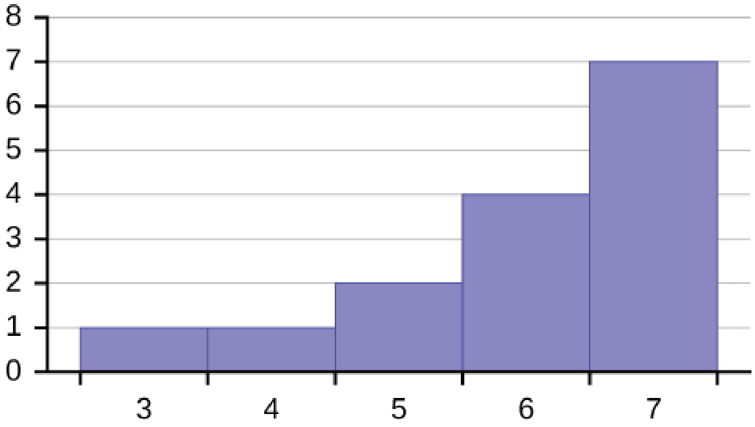
Solution:

The distribution is skewed left because it looks pulled out to the left.

Exercise:

Problem:

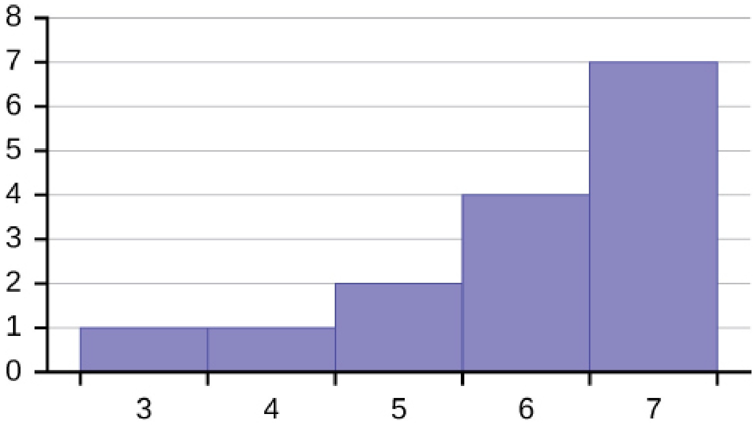
Describe the relationship between the mode and the median of this distribution.



Exercise:

Problem:

Describe the relationship between the mean and the median of this distribution.



Solution:

The mean and the median are both six.

Exercise:

Problem: The mean and median for the data are the same.

345566667777777

Is the data perfectly symmetrical? Why or why not?

Exercise:

Problem:

Which is the greatest, the mean, the mode, or the median of the data set?

11111212121212131517222222

Solution:

The mode is 12, the median is 13.5, and the mean is 15.1. The mean is the largest.

Exercise:

Problem:

Which is the least, the mean, the mode, and the median of the data set?

5656565859606264646567

Exercise:

Problem:

Of the three measures, which tends to reflect skewing the most, the mean, the mode, or the median? Why?

Solution:

The mean tends to reflect skewing the most because it is affected the most by outliers.

Exercise:

Problem:

In a perfectly symmetrical distribution, when would the mode be different from the mean and median?

Homework

Exercise:

Problem:

The median age of the U.S. population in 1980 was 30.0 years. In 1991, the median age was 33.1 years.

a. What does it mean for the median age to rise?
b. Give two reasons why the median age could rise.
c. For the median age to rise, is the actual number of children less in 1991 than it was in 1980? Why or why not?

Glossary

Skewed

used to describe data that is not symmetrical; when the right side of a graph looks “chopped off” compared the left side, we say it is “skewed to the left.” When the left side of the graph looks “chopped off” compared to the right side, we say the data is “skewed to the right.” Alternatively: when the lower values of the data are more spread out, we say the data are skewed to the left. When the greater values are more spread out, the data are skewed to the right.

Measuring the Spread of the Data

An important characteristic of any set of data is the variation in the data. In some data sets, the data values are concentrated closely near the mean; in other data sets, the data values are more widely spread out from the mean. The most common measure of variation, or spread, is the standard deviation. The **standard deviation** is a number that measures how far data values are from their mean.

The standard deviation

- provides a numerical measure of the overall amount of variation in a data set, and
- can be used to determine whether a particular data value is close to or far from the mean.

The standard deviation provides a measure of the overall variation in a data set

The standard deviation is always positive or zero. The standard deviation is small when the data are all concentrated close to the mean, exhibiting little variation or spread. The standard deviation is larger when the data values are more spread out from the mean, exhibiting more variation.

Suppose that we are studying the amount of time customers wait in line at the checkout at supermarket *A* and supermarket *B*. the average wait time at both supermarkets is five minutes. At supermarket *A*, the standard deviation for the wait time is two minutes; at supermarket *B* the standard deviation for the wait time is four minutes.

Because supermarket *B* has a higher standard deviation, we know that there is more variation in the wait times at supermarket *B*. Overall, wait times at supermarket *B* are more spread out from the average; wait times at supermarket *A* are more concentrated near the average.

The standard deviation can be used to determine whether a data value is close to or far from the mean.

Suppose that Rosa and Binh both shop at supermarket *A*. Rosa waits at the checkout counter for seven minutes and Binh waits for one minute. At supermarket *A*, the mean waiting time is five minutes and the standard deviation is two minutes. The standard deviation can be used to determine whether a data value is close to or far from the mean.

Rosa waits for seven minutes:

- Seven is two minutes longer than the average of five; two minutes is equal to one standard deviation.
- Rosa's wait time of seven minutes is **two minutes longer than the average** of five minutes.
- Rosa's wait time of seven minutes is **one standard deviation above the average** of five minutes.

Binh waits for one minute.

- One is four minutes less than the average of five; four minutes is equal to two standard deviations.
- Binh's wait time of one minute is **four minutes less than the average** of five minutes.
- Binh's wait time of one minute is **two standard deviations below the average** of five minutes.
- A data value that is two standard deviations from the average is just on the borderline for what many statisticians would consider to be far from the average. Considering data to be far from the mean if it is more than two standard deviations away is more of an approximate "rule of thumb" than a rigid rule. In general, the shape of the distribution of the data affects how much of the data is further away than two standard deviations. (You will learn more about this in later chapters.)

The number line may help you understand standard deviation. If we were to put five and seven on a number line, seven is to the right of five. We say, then, that seven is **one** standard deviation to the **right** of five because 5 + (1)(2) = 7.

If one were also part of the data set, then one is **two** standard deviations to the **left** of five because 5 + (–2)(2) = 1.

- In general, a **value = mean + (#ofSTDEV)(standard deviation)**
- where #ofSTDEVs = the number of standard deviations
- #ofSTDEV does not need to be an integer
- One is **two standard deviations less than the mean** of five because: 1 = 5 + (–2)(2).

The equation **value = mean + (#ofSTDEVs)(standard deviation)** can be expressed for a sample and for a population.

- **sample:** $x = x + (\#ofSTDEV)(s)$
- **Population:** $x = \mu + (\#ofSTDEV)(\sigma)$

The lower case letter s represents the sample standard deviation and the Greek letter σ (sigma, lower case) represents the population standard deviation.

The symbol x is the sample mean and the Greek symbol μ is the population mean.

Calculating the Standard Deviation

If x is a number, then the difference "x – mean" is called its **deviation**. In a data set, there are as many deviations as there are items in the data set. The deviations are used to calculate the standard deviation. If the numbers belong to a population, in symbols a deviation is x – μ. For sample data, in symbols a deviation is x – x.

The procedure to calculate the standard deviation depends on whether the numbers are the entire population or are data from a sample. The calculations are similar, but not identical. Therefore the symbol used to represent the standard deviation depends on whether it is calculated from a population or a sample. The lower case letter s represents the sample standard deviation and the Greek letter σ (sigma, lower case) represents the population standard deviation. If the sample has the same characteristics as the population, then s should be a good estimate of σ.

To calculate the standard deviation, we need to calculate the variance first. The **variance** is the **average of the squares of the deviations** (the x – x values for a sample, or the x – μ values for a population). The symbol σ^2 represents the population variance; the population standard deviation σ is the square root of the population variance. The symbol s^2 represents the sample variance; the sample standard deviation s is the square root of the sample variance. You can think of the standard deviation as a special average of the deviations.

If the numbers come from a census of the entire **population** and not a sample, when we calculate the average of the squared deviations to find the variance, we divide by N, the number of items in the population. If the data are from a **sample** rather than a population, when we calculate the average of the squared deviations, we divide by n **– 1**, one less than the number of items in the sample.

Formulas for the Sample Standard Deviation

- $s = \sqrt{\frac{\Sigma(x-x)^2}{n-1}}$ or $s = \sqrt{\frac{\Sigma f(x-x)^2}{n-1}}$
- For the sample standard deviation, the denominator is n **- 1**, that is the sample size MINUS 1.

Formulas for the Population Standard Deviation

- $\sigma \ = \ \sqrt{\frac{\Sigma(x-\mu)^2}{N}}$ or $\sigma \ = \ \sqrt{\frac{\Sigma f(x-\mu)^2}{N}}$
- For the population standard deviation, the denominator is N, the number of items in the population.

In these formulas, f represents the frequency with which a value appears. For example, if a value appears once, f is one. If a value appears three times in the data set or population, f is three.

Sampling Variability of a Statistic

The statistic of a sampling distribution was discussed in Descriptive Statistics: Measuring the Center of the Data. How much the statistic varies from one sample to another is known as the **sampling variability of a statistic**. You typically measure the sampling variability of a statistic by its standard error. The **standard error of the mean** is an example of a standard error. It is a special standard deviation and is known as the standard deviation of the sampling distribution of the mean. You will cover the standard error of the mean in the chapter The Central Limit Theorem (not now). The notation for the standard error of the mean is $\frac{\sigma}{\sqrt{n}}$ where σ is the standard deviation of the population and n is the size of the sample.

Note:
NOTE
In practice, USE A CALCULATOR OR COMPUTER SOFTWARE TO CALCULATE THE STANDARD DEVIATION. If you are using a TI-83, 83+, 84+ calculator, you need to select the appropriate standard deviation σ_x or s_x from the summary statistics. We will concentrate on using and interpreting the information that the standard deviation gives us. However you should study the following step-by-step example to help you understand how the standard deviation measures variation from the mean. (The calculator instructions appear at the end of this example.)

Example:
In a fifth grade class, the teacher was interested in the average age and the sample standard deviation of the ages of her students. The following data are the ages for a SAMPLE of n = 20 fifth grade students. The ages are rounded to the nearest half year:
9; 9.5; 9.5; 10; 10; 10; 10; 10.5; 10.5; 10.5; 10.5; 11; 11; 11; 11; 11; 11; 11.5; 11.5; 11.5;

Equation:

$$x = \frac{9 + 9.5(2) + 10(4) + 10.5(4) + 11(6) + 11.5(3)}{20} = 10.525$$

The average age is 10.53 years, rounded to two places.
The variance may be calculated by using a table. Then the standard deviation is calculated by taking the square root of the variance. We will explain the parts of the table after calculating s.

Data	**Freq.**	**Deviations**	***Deviations***^{2}	**(Freq.) (*Deviations*^{2})**
x	f	$(x – x)$	$(x – x)^2$	$(f)(x – x)^2$
9	1	9 – 10.525 = –1.525	$(–1.525)^2$ = 2.325625	1 × 2.325625 = 2.325625
9.5	2	9.5 – 10.525 = –1.025	$(–1.025)^2$ = 1.050625	2 × 1.050625 = 2.101250
10	4	10 – 10.525 = –0.525	$(–0.525)^2$ = 0.275625	4 × 0.275625 = 1.1025
10.5	4	10.5 – 10.525 = –0.025	$(–0.025)^2$ = 0.000625	4 × 0.000625 = 0.0025
11	6	11 – 10.525 = 0.475	$(0.475)^2$ = 0.225625	6 × 0.225625 = 1.35375
11.5	3	11.5 – 10.525 = 0.975	$(0.975)^2$ = 0.950625	3 × 0.950625 = 2.851875

Data	**Freq.**	**Deviations**	***Deviations*^{2}**	**(Freq.)(*Deviations*^{2})**
				The total is 9.7375

The sample variance, s^2, is equal to the sum of the last column (9.7375) divided by the total number of data values minus one (20 – 1):

$s^2 = \frac{9.7375}{20-1} = 0.5125$

The **sample standard deviation** s is equal to the square root of the sample variance:

$s = \sqrt{0.5125} = 0.715891$, which is rounded to two decimal places, s = 0.72.

Typically, you do the calculation for the standard deviation on your calculator or computer. The intermediate results are not rounded. This is done for accuracy.

Exercise:

Problem:

- For the following problems, recall that **value = mean + (#ofSTDEVs)(standard deviation)**. Verify the mean and standard deviation or a calculator or computer.
- For a sample: x = x + (#ofSTDEVs)(s)
- For a population: x= μ + (#ofSTDEVs)(σ)
- For this example, use x = x + (#ofSTDEVs)(s) because the data is from a sample

a. Verify the mean and standard deviation on your calculator or computer.
b. Find the value that is one standard deviation above the mean. Find (x + 1s).
c. Find the value that is two standard deviations below the mean. Find (x – 2s).
d. Find the values that are 1.5 standard deviations **from** (below and above) the mean.

Solution:

a. **Note:**

   - Clear lists L1 and L2. Press STAT 4:ClrList. Enter 2nd 1 for L1, the comma (,), and 2nd 2 for L2.
   - Enter data into the list editor. Press STAT 1:EDIT. If necessary, clear the lists by arrowing up into the name. Press CLEAR and arrow down.

    - Put the data values (9, 9.5, 10, 10.5, 11, 11.5) into list L1 and the frequencies (1, 2, 4, 4, 6, 3) into list L2. Use the arrow keys to move around.
    - Press STAT and arrow to CALC. Press 1:1-VarStats and enter L1 (2nd 1), L2 (2nd 2). Do not forget the comma. Press ENTER.
    - x = 10.525
    - Use Sx because this is sample data (not a population): Sx=0.715891

b. (x + 1s) = 10.53 + (1)(0.72) = 11.25
c. (x – 2s) = 10.53 – (2)(0.72) = 9.09

d.
- (x – 1.5s) = 10.53 – (1.5)(0.72) = 9.45
- (x + 1.5s) = 10.53 + (1.5)(0.72) = 11.61

Note:
Try It
Exercise:

Problem: On a baseball team, the ages of each of the players are as follows:

21; 21; 22; 23; 24; 24; 25; 25; 28; 29; 29; 31; 32; 33; 33; 34; 35; 36; 36; 36; 36; 38; 38; 38; 40

Use your calculator or computer to find the mean and standard deviation. Then find the value that is two standard deviations above the mean.

Solution:

μ = 30.68

s = 6.09
(x + 2s) = 30.68 + (2)(6.09) = 42.86.

Explanation of the standard deviation calculation shown in the table

The deviations show how spread out the data are about the mean. The data value 11.5 is farther from the mean than is the data value 11 which is indicated by the deviations 0.97 and 0.47. A positive deviation occurs when the data value is greater than the mean, whereas a negative deviation occurs when the data value is less than the mean. The deviation is –1.525 for the data value nine. **If you add the deviations, the sum is always zero**. (For [link], there are n = 20 deviations.) So you cannot simply add the deviations to get the spread of the data. By squaring the deviations, you make them positive numbers, and the sum will also be positive. The variance, then, is the average squared deviation.

The variance is a squared measure and does not have the same units as the data. Taking the square root solves the problem. The standard deviation measures the spread in the same units as the data.

Notice that instead of dividing by n = 20, the calculation divided by n – 1 = 20 – 1 = 19 because the data is a sample. For the **sample** variance, we divide by the sample size minus one (n – 1). Why not divide by n? The answer has to do with the population variance. **The sample variance is an estimate of the population variance.** Based on the theoretical mathematics that lies behind these calculations, dividing by (n – 1) gives a better estimate of the population variance.

Note:
NOTE
Your concentration should be on what the standard deviation tells us about the data. The standard deviation is a number which measures how far the data are spread from the mean. Let a calculator or computer do the arithmetic.

The standard deviation, s or σ, is either zero or larger than zero. When the standard deviation is zero, there is no spread; that is, the all the data values are equal to each other. The standard deviation is small when the data are all concentrated close to the mean, and is larger when the data values show more variation from the mean. When the standard deviation is a lot larger than zero, the data values are very spread out about the mean; outliers can make s or σ very large.

The standard deviation, when first presented, can seem unclear. By graphing your data, you can get a better "feel" for the deviations and the standard deviation. You will find that in symmetrical distributions, the standard deviation can be very helpful but in skewed distributions, the standard deviation may not be much help. The reason is that the two sides of a skewed distribution have different spreads. In a skewed distribution, it is better to look at the first quartile, the median, the third quartile, the smallest value, and the largest value. Because numbers can be confusing, **always graph your data**. Display your data in a histogram or a box plot.

Example:
Exercise:

Problem:

Use the following data (first exam scores) from Susan Dean's spring pre-calculus class:

33; 42; 49; 49; 53; 55; 55; 61; 63; 67; 68; 68; 69; 69; 72; 73; 74; 78; 80; 83; 88; 88; 88; 90; 92; 94; 94; 94; 94; 96; 100

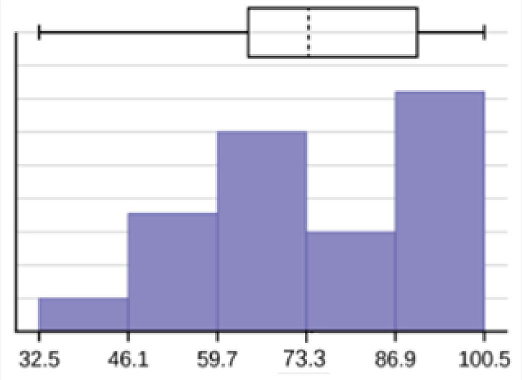
a. Create a chart containing the data, frequencies, relative frequencies, and cumulative relative frequencies to three decimal places.
b. Calculate the following to one decimal place using a TI-83+ or TI-84 calculator:
    i. The sample mean
    ii. The sample standard deviation
    iii. The median
    iv. The first quartile
    v. The third quartile
    vi. *IQR*
c. Construct a box plot and a histogram on the same set of axes. Make comments about the box plot, the histogram, and the chart.

Solution:

a. See [link]

b.
    i. The sample mean = 73.5
    ii. The sample standard deviation = 17.9
    iii. The median = 73
    iv. The first quartile = 61
    v. The third quartile = 90
    vi. *IQR* = 90 – 61 = 29

c. The x-axis goes from 32.5 to 100.5; y-axis goes from –2.4 to 15 for the histogram. The number of intervals is five, so the width of an interval is (100.5 – 32.5) divided by five, is equal to 13.6. Endpoints of the intervals are as follows: the starting point is 32.5, 32.5 + 13.6 = 46.1, 46.1 + 13.6 = 59.7, 59.7 + 13.6 = 73.3, 73.3 + 13.6 = 86.9, 86.9 + 13.6 = 100.5 = the ending value; No data values fall on an interval boundary.



The long left whisker in the box plot is reflected in the left side of the histogram. The spread of the exam scores in the lower 50% is greater (73 – 33 = 40) than the spread in the upper 50% (100 – 73 = 27). The histogram, box plot, and chart all reflect this. There are a substantial number of A and B grades (80s, 90s, and 100). The histogram clearly shows this. The box plot shows us that the middle 50% of the exam scores (*IQR* = 29) are Ds, Cs, and Bs. The box plot also shows us that the lower 25% of the exam scores are Ds and Fs.

Data	**Frequency**	**Relative Frequency**	**Cumulative Relative Frequency**
33	1	0.032	0.032
42	1	0.032	0.064
49	2	0.065	0.129
53	1	0.032	0.161
55	2	0.065	0.226

Data	Frequency	Relative Frequency	Cumulative Relative Frequency
61	1	0.032	0.258
63	1	0.032	0.29
67	1	0.032	0.322
68	2	0.065	0.387
69	2	0.065	0.452
72	1	0.032	0.484
73	1	0.032	0.516
74	1	0.032	0.548
78	1	0.032	0.580
80	1	0.032	0.612
83	1	0.032	0.644
88	3	0.097	0.741
90	1	0.032	0.773
92	1	0.032	0.805
94	4	0.129	0.934
96	1	0.032	0.966
100	1	0.032	0.998 (Why isn't this value 1?)

Note:
Try It
Exercise:

Problem:

The following data show the different types of pet food stores in the area carry.
6; 6; 6; 6; 7; 7; 7; 7; 7; 8; 9; 9; 9; 9; 10; 10; 10; 10; 10; 11; 11; 11; 11; 12; 12; 12; 12; 12; 12;
Calculate the sample mean and the sample standard deviation to one decimal place using a TI-83+ or TI-84 calculator.

Solution:

μ = 9.3

s = 2.2

Standard deviation of Grouped Frequency Tables

Recall that for grouped data we do not know individual data values, so we cannot describe the typical value of the data with precision. In other words, we cannot find the exact mean, median, or mode. We can, however, determine the best estimate of the measures of center by finding the mean of the grouped data with the formula:

$$Mean\ of\ Frequency\ Table = \frac{\sum fm}{\sum f}$$

where $f =$ interval frequencies and m = interval midpoints.

Just as we could not find the exact mean, neither can we find the exact standard deviation. Remember that standard deviation describes numerically the expected deviation a data value has from the mean. In simple English, the standard deviation allows us to compare how “unusual” individual data is compared to the mean.

Example:
Find the standard deviation for the data in the following table.

Class	**Frequency, *f***	**Midpoint, *m***	m^2	x^2	fm^2	**Standard Deviation**
0–2	1	1	1	7.58	1	3.5
3–5	6	4	16	7.58	96	3.5
6–8	10	7	49	7.58	490	3.5
9–11	7	10	100	7.58	700	3.5
12–14	0	13	169	7.58	0	3.5
15–17	2	16	256	7.58	512	3.5

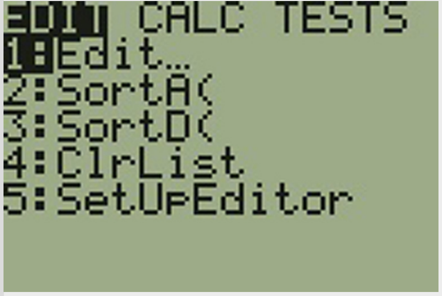
For this data set, we have the mean, x = 7.58 and the standard deviation, s_x = 3.5. This means that a randomly selected data value would be expected to be 3.5 units from the mean. If we look at the first class, we see that the class midpoint is equal to one. This is almost two full standard deviations from the mean since 7.58 – 3.5 – 3.5 = 0.58. While the formula for calculating the standard deviation is not complicated, $s_x = \sqrt{\frac{f(m-x)^2}{n-1}}$ where s_x = sample standard deviation, x = sample mean, the calculations are tedious. It is usually best to use technology when performing the calculations.

Note:
Try It
Find the standard deviation for the data from the previous example

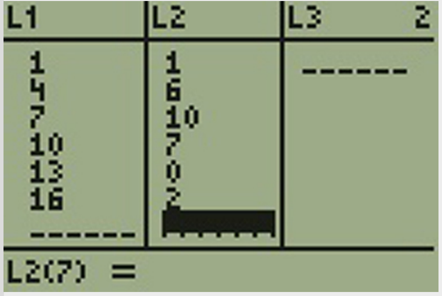
Class	**Frequency, *f***
0–2	1

Class	**Frequency, *f***
3–5	6
6–8	10
9–11	7
12–14	0
15–17	2

First, press the **STAT** key and select **1:Edit**


EDIT CALC TESTS
1:Edit…
2:SortA(
3:SortD(
4:ClrList
5:SetUpEditor

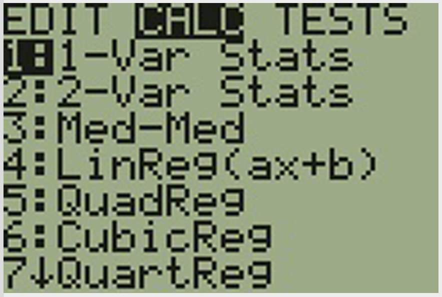
Input the midpoint values into **L1** and the frequencies into **L2**



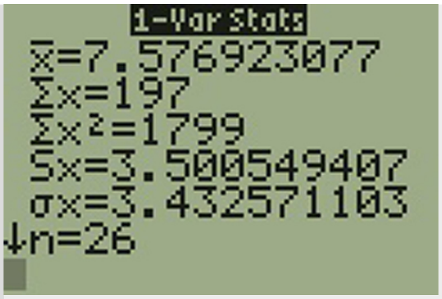
L1	L2	L3 2
1	1	------
4	6	
7	10	
10	7	
13	0	
16	2	

L2(7) =

Select **STAT**, **CALC**, and **1: 1-Var Stats**


EDIT CALC TESTS
1:1-Var Stats
2:2-Var Stats
3:Med-Med
4:LinReg(ax+b)
5:QuadReg
6:CubicReg
7↓QuartReg

Select **2nd** then **1** then , **2nd** then **2 Enter**



```
1-Var Stats
 x̄=7.576923077
 Σx=197
 Σx²=1799
 Sx=3.500549407
 σx=3.432571103
↓n=26
```

You will see displayed both a population standard deviation, σ_x, and the sample standard deviation, s_x.

Comparing Values from Different Data Sets

The standard deviation is useful when comparing data values that come from different data sets. If the data sets have different means and standard deviations, then comparing the data values directly can be misleading.

- For each data value, calculate how many standard deviations away from its mean the value is.
- Use the formula: value = mean + (#ofSTDEVs)(standard deviation); solve for #ofSTDEVs.
- $\#ofSTDEVs = \frac{\text{value – mean}}{\text{standard deviation}}$
- Compare the results of this calculation.

#ofSTDEVs is often called a "z-score"; we can use the symbol z. In symbols, the formulas become:

Sample	$x = x + zs$	$z = \frac{x - x}{s}$
Population	$x = \mu + z\sigma$	$z = \frac{x - \mu}{\sigma}$

Example:
Exercise:

Problem:

Two students, John and Ali, from different high schools, wanted to find out who had the highest GPA when compared to his school. Which student had the highest GPA when compared to his school?

Student	GPA	School Mean GPA	School Standard Deviation
John	2.85	3.0	0.7
Ali	77	80	10

Solution:

For each student, determine how many standard deviations (#ofSTDEVs) his GPA is away from the average, for his school. Pay careful attention to signs when comparing and interpreting the answer.

$$z = \#\text{ of STDEVs} = \frac{\text{value} - \text{mean}}{\text{standard deviation}} = \frac{x - \mu}{\sigma}$$

For John, $z = \#ofSTDEVs = \frac{2.85 - 3.0}{0.7} = -0.21$

For Ali, $z = \#ofSTDEVs = \frac{77 - 80}{10} = -0.3$

John has the better GPA when compared to his school because his GPA is 0.21 standard deviations **below** his school's mean while Ali's GPA is 0.3 standard deviations **below** his school's mean.

John's z-score of –0.21 is higher than Ali's z-score of –0.3. For GPA, higher values are better, so we conclude that John has the better GPA when compared to his school.

Note:
Try It
Exercise:

Problem:

Two swimmers, Angie and Beth, from different teams, wanted to find out who had the fastest time for the 50 meter freestyle when compared to her team. Which swimmer had the fastest time when compared to her team?

Swimmer	Time (seconds)	Team Mean Time	Team Standard Deviation
Angie	26.2	27.2	0.8
Beth	27.3	30.1	1.4

Solution:

For Angie: $z = \frac{26.2 - 27.2}{0.8} = -1.25$

For Beth: $z = \frac{27.3 - 30.1}{1.4} = -2$

The following lists give a few facts that provide a little more insight into what the standard deviation tells us about the distribution of the data.

For ANY data set, no matter what the distribution of the data is:

- At least 75% of the data is within two standard deviations of the mean.
- At least 89% of the data is within three standard deviations of the mean.
- At least 95% of the data is within 4.5 standard deviations of the mean.
- This is known as **Chebyshev's Rule**.

For data having a distribution that is BELL-SHAPED and SYMMETRIC:

- Approximately 68% of the data is within one standard deviation of the mean.
- Approximately 95% of the data is within two standard deviations of the mean.
- More than 99.7% of the data is within three standard deviations of the mean.
- This is known as the **Empirical Rule**.
- It is important to note that this rule only applies when the shape of the distribution of the data is bell-shaped and symmetric. We will learn more about this when studying the "Normal" or "Gaussian" probability distribution in later chapters.

References

Data from Microsoft Bookshelf.

King, Bill.“Graphically Speaking.” Institutional Research, Lake Tahoe Community College. Available online at http://www.ltcc.edu/web/about/institutional-research (accessed April 3, 2013).

Section Review

The standard deviation can help you calculate the spread of data. There are different equations to use if are calculating the standard deviation of a sample or of a population.

- The Standard Deviation allows us to compare individual data or classes to the data set mean numerically.
- $s = \sqrt{\frac{\sum (x-x)^2}{n-1}}$ or $s = \sqrt{\frac{\sum f(x-x)^2}{n-1}}$ is the formula for calculating the standard deviation of a sample. To calculate the standard deviation of a population, we would use the population mean, μ, and the formula $\sigma = \sqrt{\frac{\sum (x-\mu)^2}{N}}$ or $\sigma = \sqrt{\frac{\sum f(x-\mu)^2}{N}}$.

Use the following information to answer the next two exercises: The following data are the distances between 20 retail stores and a large distribution center. The distances are in miles.
29; 37; 38; 40; 58; 67; 68; 69; 76; 86; 87; 95; 96; 96; 99; 106; 112; 127; 145; 150
Exercise:

Problem:

Use a graphing calculator or computer to find the standard deviation and round to the nearest tenth.

Solution:

$s = 34.5$

Exercise:

Problem: Find the value that is one standard deviation below the mean.

Exercise:

Problem:

Two baseball players, Fredo and Karl, on different teams wanted to find out who had the higher batting average when compared to his team. Which baseball player had the higher batting average when compared to his team?

Baseball Player	**Batting Average**	**Team Batting Average**	**Team Standard Deviation**
Fredo	0.158	0.166	0.012
Karl	0.177	0.189	0.015

Solution:

For Fredo: $z = \frac{0.158 - 0.166}{0.012} = -0.67$

For Karl: $z = \frac{0.177 - 0.189}{0.015} = -0.8$

Fredo’s z-score of –0.67 is higher than Karl’s z-score of –0.8. For batting average, higher values are better, so Fredo has a better batting average compared to his team.

Exercise:

Problem: Use [link] to find the value that is three standard deviations:

- **a**above the mean

- **b**below the mean

Find the standard deviation for the following frequency tables using the formula. Check the calculations with the TI 83/84.

Exercise:

Problem:

Find the standard deviation for the following frequency tables using the formula. Check the calculations with the TI 83/84.

a.

Grade	Frequency
49.5–59.5	2
59.5–69.5	3
69.5–79.5	8
79.5–89.5	12
89.5–99.5	5

b.

Daily Low Temperature	Frequency
49.5–59.5	53
59.5–69.5	32
69.5–79.5	15

Daily Low Temperature	Frequency
79.5–89.5	1
89.5–99.5	0

c.

Points per Game	Frequency
49.5–59.5	14
59.5–69.5	32
69.5–79.5	15
79.5–89.5	23
89.5–99.5	2

Solution:

a. $s_x = \sqrt{\frac{\sum fm^2}{n} - x^2} = \sqrt{\frac{193157.45}{30} - 79.5^2} = 10.88$

b. $s_x = \sqrt{\frac{\sum fm^2}{n} - x^2} = \sqrt{\frac{380945.3}{101} - 60.94^2} = 7.62$

c. $s_x = \sqrt{\frac{\sum fm^2}{n} - x^2} = \sqrt{\frac{440051.5}{86} - 70.66^2} = 11.14$

Homework

Use the following information to answer the next nine exercises: The population parameters below describe the full-time equivalent number of students (FTES) each year at Lake Tahoe Community College from 1976–1977 through 2004–2005.

- μ = 1000 FTES
- median = 1,014 FTES
- σ = 474 FTES
- first quartile = 528.5 FTES
- third quartile = 1,447.5 FTES
- n = 29 years

Exercise:

Problem:

A sample of 11 years is taken. About how many are expected to have a FTES of 1014 or above? Explain how you determined your answer.

Solution:

The median value is the middle value in the ordered list of data values. The median value of a set of 11 will be the 6th number in order. Six years will have totals at or below the median.

Exercise:

Problem: 75% of all years have an FTES:

a. at or below: _____
b. at or above: _____

Exercise:

Problem: The population standard deviation = _____

Solution:

474 FTES

Exercise:

Problem:

What percent of the FTES were from 528.5 to 1447.5? How do you know?

Exercise:

Problem: What is the *IQR*? What does the *IQR* represent?

Solution:

919

Exercise:

Problem: How many standard deviations away from the mean is the median?

Additional Information: The population FTES for 2005–2006 through 2010–2011 was given in an updated report. The data are reported here.

Year	2005–06	2006–07	2007–08	2008–09	2009–10	2010–11
Total FTES	1,585	1,690	1,735	1,935	2,021	1,890

Exercise:

Problem:

Calculate the mean, median, standard deviation, the first quartile, the third quartile and the *IQR*. Round to one decimal place.

Solution:

- mean = 1,809.3
- median = 1,812.5
- standard deviation = 151.2
- first quartile = 1,690
- third quartile = 1,935
- *IQR* = 245

Exercise:

Problem:

Construct a box plot for the FTES for 2005–2006 through 2010–2011 and a box plot for the FTES for 1976–1977 through 2004–2005.

Exercise:

Problem:

Compare the *IQR* for the FTES for 1976–77 through 2004–2005 with the *IQR* for the FTES for 2005-2006 through 2010–2011. Why do you suppose the *IQR*s are so different?

Solution:

Hint: Think about the number of years covered by each time period and what happened to higher education during those periods.

Exercise:

Problem:

Three students were applying to the same graduate school. They came from schools with different grading systems. Which student had the best GPA when compared to other students at his school? Explain how you determined your answer.

Student	GPA	School Average GPA	School Standard Deviation
Thuy	2.7	3.2	0.8
Vichet	87	75	20
Kamala	8.6	8	0.4

Exercise:

Problem:

A music school has budgeted to purchase three musical instruments. They plan to purchase a piano costing $3,000, a guitar costing $550, and a drum set costing $600. The mean cost for a piano is $4,000 with a standard deviation of $2,500. The mean cost for a guitar is $500 with a standard deviation of $200. The mean cost for drums is $700 with a standard deviation of $100. Which cost is the lowest, when compared to other instruments of the same type? Which cost is the highest when compared to other instruments of the same type. Justify your answer.

Solution:

For pianos, the cost of the piano is 0.4 standard deviations BELOW the mean. For guitars, the cost of the guitar is 0.25 standard deviations ABOVE the mean. For drums, the cost of the drum set is 1.0 standard deviations BELOW the mean. Of the three, the drums cost the lowest in comparison to the cost of other instruments of the same type. The guitar costs the most in comparison to the cost of other instruments of the same type.

Exercise:

Problem:

An elementary school class ran one mile with a mean of 11 minutes and a standard deviation of three minutes. Rachel, a student in the class, ran one mile in eight minutes. A junior high school class ran one mile with a mean of nine minutes and a standard deviation of two minutes. Kenji, a student in the class, ran 1 mile in 8.5 minutes. A high school class ran one mile with a mean of seven minutes and a standard deviation of four minutes. Nedda, a student in the class, ran one mile in eight minutes.

a. Why is Kenji considered a better runner than Nedda, even though Nedda ran faster than he?
b. Who is the fastest runner with respect to his or her class? Explain why.

Exercise:

Problem:

The most obese countries in the world have obesity rates that range from 11.4% to 74.6%. This data is summarized in Table 14.

Percent of Population Obese	Number of Countries
11.4–20.45	29
20.45–29.45	13
29.45–38.45	4
38.45–47.45	0
47.45–56.45	2
56.45–65.45	1
65.45–74.45	0
74.45–83.45	1

What is the best estimate of the average obesity percentage for these countries? What is the standard deviation for the listed obesity rates? The United States has an average obesity rate of 33.9%. Is this rate above average or below? How “unusual” is the United States’ obesity rate compared to the average rate? Explain.

Solution:

- $x = 23.32$
- Using the TI 83/84, we obtain a standard deviation of: $s_x = 12.95$.
- The obesity rate of the United States is 10.58% higher than the average obesity rate.
- Since the standard deviation is 12.95, we see that 23.32 + 12.95 = 36.27 is the obesity percentage that is one standard deviation from the mean. The United States obesity rate is slightly less than one standard deviation from the mean. Therefore, we can assume that the United States, while 34% obese, does not have an unusually high percentage of obese people.

Exercise:

Problem:

[link] gives the percent of children under five considered to be underweight.

Percent of Underweight Children	Number of Countries
16–21.45	23
21.45–26.9	4
26.9–32.35	9
32.35–37.8	7
37.8–43.25	6
43.25–48.7	1

What is the best estimate for the mean percentage of underweight children? What is the standard deviation? Which interval(s) could be considered unusual? Explain.

Bringing It Together

Exercise:

Problem:

Twenty-five randomly selected students were asked the number of movies they watched the previous week. The results are as follows:

# of movies	Frequency
0	5
1	9
2	6
3	4

# of movies	Frequency
4	1

a. Find the sample mean x.
b. Find the approximate sample standard deviation, s.

Solution:

a. 1.48
b. 1.12

Exercise:

Problem:

Forty randomly selected students were asked the number of pairs of sneakers they owned. Let X = the number of pairs of sneakers owned. The results are as follows:

X	**Frequency**
1	2
2	5
3	8
4	12
5	12
6	0
7	1

a. Find the sample mean x

b. Find the sample standard deviation, s
c. Construct a histogram of the data.
d. Complete the columns of the chart.
e. Find the first quartile.
f. Find the median.
g. Find the third quartile.
h. Construct a box plot of the data.
i. What percent of the students owned at least five pairs?
j. Find the 40th percentile.
k. Find the 90th percentile.
l. Construct a line graph of the data
m. Construct a stemplot of the data

Exercise:

Problem:

Following are the published weights (in pounds) of all of the team members of the San Francisco 49ers from a previous year.

177; 205; 210; 210; 232; 205; 185; 185; 178; 210; 206; 212; 184; 174; 185; 242; 188; 212; 215; 247; 241; 223; 220; 260; 245; 259; 278; 270; 280; 295; 275; 285; 290; 272; 273; 280; 285; 286; 200; 215; 185; 230; 250; 241; 190; 260; 250; 302; 265; 290; 276; 228; 265

a. Organize the data from smallest to largest value.
b. Find the median.
c. Find the first quartile.
d. Find the third quartile.
e. Construct a box plot of the data.
f. The middle 50% of the weights are from _______ to _______.
g. If our population were all professional football players, would the above data be a sample of weights or the population of weights? Why?
h. If our population included every team member who ever played for the San Francisco 49ers, would the above data be a sample of weights or the population of weights? Why?
i. Assume the population was the San Francisco 49ers. Find:

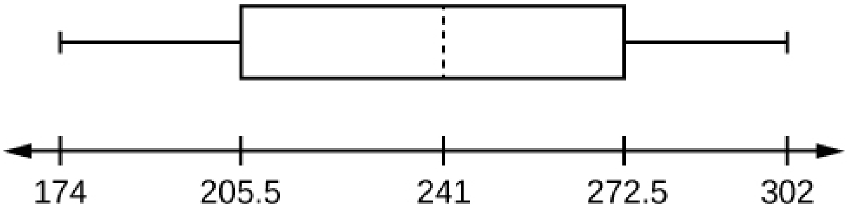
    i. the population mean, μ.
    ii. the population standard deviation, σ.
    iii. the weight that is two standard deviations below the mean.
    iv. When Steve Young, quarterback, played football, he weighed 205 pounds. How many standard deviations above or below the mean was he?

j. That same year, the mean weight for the Dallas Cowboys was 240.08 pounds with a standard deviation of 44.38 pounds. Emmit Smith weighed in at 209 pounds. With respect to his team, who was lighter, Smith or Young? How did you determine your answer?

Solution:

a. 174; 177; 178; 184; 185; 185; 185; 185; 188; 190; 200; 205; 205; 206; 210; 210; 210; 212; 212; 215; 215; 220; 223; 228; 230; 232; 241; 241; 242; 245; 247; 250; 250; 259; 260; 260; 265; 265; 270; 272; 273; 275; 276; 278; 280; 280; 285; 285; 286; 290; 290; 295; 302
b. 241
c. 205.5
d. 272.5

e.



f. 205.5, 272.5
g. sample
h. population

i.
    i. 236.34
    ii. 37.50
    iii. 161.34
    iv. 0.84 std. dev. below the mean

j. Young

Exercise:

Problem:

One hundred teachers attended a seminar on mathematical problem solving. The attitudes of a representative sample of 12 of the teachers were measured before and after the seminar. A positive number for change in attitude indicates that a teacher's attitude toward math became more positive. The 12 change scores are as follows:

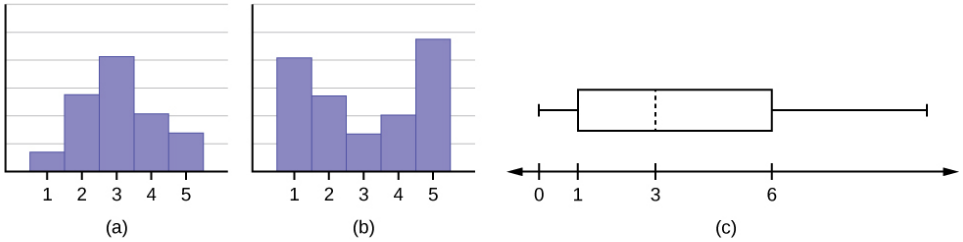
3 8–12 05–31–16 5–2

a. What is the mean change score?
b. What is the standard deviation for this population?
c. What is the median change score?
d. Find the change score that is 2.2 standard deviations below the mean.

Exercise:

Problem:

Refer to [link] determine which of the following are true and which are false. Explain your solution to each part in complete sentences.



a. The medians for all three graphs are the same.
b. We cannot determine if any of the means for the three graphs is different.
c. The standard deviation for graph b is larger than the standard deviation for graph a.
d. We cannot determine if any of the third quartiles for the three graphs is different.

Solution:

a. True
b. True
c. True
d. False

Exercise:

Problem:

In a recent issue of the *IEEE Spectrum*, 84 engineering conferences were announced. Four conferences lasted two days. Thirty-six lasted three days. Eighteen lasted four days. Nineteen lasted five days. Four lasted six days. One lasted seven days. One lasted eight days. One lasted nine days. Let X = the length (in days) of an engineering conference.

a. Organize the data in a chart.

b. Find the median, the first quartile, and the third quartile.
c. Find the 65th percentile.
d. Find the 10th percentile.
e. Construct a box plot of the data.
f. The middle 50% of the conferences last from _______ days to _______ days.
g. Calculate the sample mean of days of engineering conferences.
h. Calculate the sample standard deviation of days of engineering conferences.
i. Find the mode.
j. If you were planning an engineering conference, which would you choose as the length of the conference: mean; median; or mode? Explain why you made that choice.
k. Give two reasons why you think that three to five days seem to be popular lengths of engineering conferences.

Exercise:

Problem:

A survey of enrollment at 35 community colleges across the United States yielded the following figures:

6414; 1550; 2109; 9350; 21828; 4300; 5944; 5722; 2825; 2044; 5481; 5200; 5853; 2750; 10012; 6357; 27000; 9414; 7681; 3200; 17500; 9200; 7380; 18314; 6557; 13713; 17768; 7493; 2771; 2861; 1263; 7285; 28165; 5080; 11622

a. Organize the data into a chart with five intervals of equal width. Label the two columns "Enrollment" and "Frequency."
b. Construct a histogram of the data.
c. If you were to build a new community college, which piece of information would be more valuable: the mode or the mean?
d. Calculate the sample mean.
e. Calculate the sample standard deviation.
f. A school with an enrollment of 8000 would be how many standard deviations away from the mean?

Solution:

a.

Enrollment	**Frequency**
1000-5000	10
5000-10000	16
10000-15000	3
15000-20000	3
20000-25000	1
25000-30000	2

b. Check student’s solution.
c. mode
d. 8628.74
e. 6943.88
f. –0.09

Use the following information to answer the next two exercises. X = the number of days per week that 100 clients use a particular exercise facility.

x	**Frequency**
0	3
1	12
2	33
3	28
4	11

x	**Frequency**
5	9
6	4

Exercise:

Problem: The 80th percentile is _____

a. 5
b. 80
c. 3
d. 4

Exercise:

Problem:

The number that is 1.5 standard deviations BELOW the mean is approximately _____

a. 0.7
b. 4.8
c. –2.8
d. Cannot be determined

Solution:

a

Exercise:

Problem:

Suppose that a publisher conducted a survey asking adult consumers the number of fiction paperback books they had purchased in the previous month. The results are summarized in the [link].

# of books	**Freq.**	**Rel. Freq.**
0	18	
1	24	
2	24	
3	22	
4	15	
5	10	
7	5	
9	1	

a. Are there any outliers in the data? Use an appropriate numerical test involving the *IQR* to identify outliers, if any, and clearly state your conclusion.
b. If a data value is identified as an outlier, what should be done about it?
c. Are any data values further than two standard deviations away from the mean? In some situations, statisticians may use this criteria to identify data values that are unusual, compared to the other data values. (Note that this criteria is most appropriate to use for data that is mound-shaped and symmetric, rather than for skewed data.)
d. Do parts a and c of this problem give the same answer?
e. Examine the shape of the data. Which part, a or c, of this question gives a more appropriate result for this data?
f. Based on the shape of the data which is the most appropriate measure of center for this data: mean, median or mode?

Glossary

Standard Deviation
a number that is equal to the square root of the variance and measures how far data values are from their mean; notation: s for sample standard deviation and σ for population standard deviation.

Variance

mean of the squared deviations from the mean, or the square of the standard deviation; for a set of data, a deviation can be represented as $x – x$ where x is a value of the data and x is the sample mean. The sample variance is equal to the sum of the squares of the deviations divided by the difference of the sample size and one.

# Probability Topics: Introduction
class="introduction"

Meteor showers are rare, but the probability of them occurring can be calculated. (credit: Navicore/flickr)

Note:
Chapter Objectives
By the end of this chapter, the student should be able to:

- Understand and use the terminology of probability.
- Determine whether two events are mutually exclusive and whether two events are independent.

- Calculate probabilities using the Addition Rules and Multiplication Rules.
- Construct and interpret Contingency Tables.
- Construct and interpret Venn Diagrams.(optional)
- Construct and interpret Tree Diagrams.(optional)

It is often necessary to "guess" about the outcome of an event in order to make a decision. Politicians study polls to guess their likelihood of winning an election. Teachers choose a particular course of study based on what they think students can comprehend. Doctors choose the treatments needed for various diseases based on their assessment of likely results. You may have visited a casino where people play games chosen because of the belief that the likelihood of winning is good. You may have chosen your course of study based on the probable availability of jobs.

You have, more than likely, used probability. In fact, you probably have an intuitive sense of probability. Probability deals with the chance of an event occurring. Whenever you weigh the odds of whether or not to do your homework or to study for an exam, you are using probability. In this chapter, you will learn how to solve probability problems using a systematic approach.

Note:
Collaborative Exercise
Your instructor will survey your class. Count the number of students in the class today.

- Raise your hand if you have any change in your pocket or purse. Record the number of raised hands.
- Raise your hand if you rode a bus within the past month. Record the number of raised hands.

- Raise your hand if you answered "yes" to BOTH of the first two questions. Record the number of raised hands.

Use the class data as estimates of the following probabilities. P(change) means the probability that a randomly chosen person in your class has change in his/her pocket or purse. P(bus) means the probability that a randomly chosen person in your class rode a bus within the last month and so on. Discuss your answers.

- Find P(change).
- Find P(bus).
- Find P(change AND bus). Find the probability that a randomly chosen student in your class has change in his/her pocket or purse and rode a bus within the last month.
- Find P(change|bus). Find the probability that a randomly chosen student has change given that he or she rode a bus within the last month. Count all the students that rode a bus. From the group of students who rode a bus, count those who have change. The probability is equal to those who have change and rode a bus divided by those who rode a bus.

Terminology

Probability is a measure that is associated with how certain we are of outcomes of a particular experiment or activity. An **experiment** is a planned operation carried out under controlled conditions. If the result is not predetermined, then the experiment is said to be a **chance** experiment. Flipping one fair coin twice is an example of an experiment.

A result of an experiment is called an **outcome**. The **sample space** of an experiment is the set of all possible outcomes. Three ways to represent a sample space are: to list the possible outcomes, to create a tree diagram, or to create a Venn diagram. The uppercase letter S is used to denote the sample space. For example, if you flip one fair coin, $S = \{H, T\}$ where H = heads and T = tails are the outcomes.

An **event** is any combination of outcomes. Upper case letters like A and B represent events. For example, if the experiment is to flip one fair coin, event A might be getting at most one head. The probability of an event A is written $P(A)$.

The **probability** of any outcome is the **long-term relative frequency** of that outcome. **Probabilities are between zero and one, inclusive** (that is, zero and one and all numbers between these values). $P(A) = 0$ means the event A can never happen. $P(A) = 1$ means the event A always happens. $P(A) = 0.5$ means the event A is equally likely to occur or not to occur. For example, if you flip one fair coin repeatedly (from 20 to 2,000 to 20,000 times) the relative frequency of heads approaches 0.5 (the probability of heads).

Equally likely means that each outcome of an experiment occurs with equal probability. For example, if you toss a **fair**, six-sided die, each face (1, 2, 3, 4, 5, or 6) is as likely to occur as any other face. If you toss a fair coin, a Head (H) and a Tail (T) are equally likely to occur. If you randomly guess the answer to a true/false question on an exam, you are equally likely to select a correct answer or an incorrect answer.

To calculate the probability of an event *A* when all outcomes in the sample space are equally likely, count the number of outcomes for event *A* and divide by the total number of outcomes in the sample space. For example, if you toss a fair dime and a fair nickel, the sample space is {*HH*, *TH*, *HT*, *TT*} where *T* = tails and *H* = heads. The sample space has four outcomes. *A* = getting one head. There are two outcomes that meet this condition {*HT*, *TH*}, so $P(A)$ = $\frac{2}{4}$ = 0.5.

Suppose you roll one fair six-sided die, with the numbers {1, 2, 3, 4, 5, 6} on its faces. Let event *E* = rolling a number that is at least five. There are two outcomes {5, 6}. $P(E)$ = $\frac{2}{6}$. If you were to roll the die only a few times, you would not be surprised if your observed results did not match the probability. If you were to roll the die a very large number of times, you would expect that, overall, $\frac{2}{6}$ of the rolls would result in an outcome of "at least five". You would not expect exactly $\frac{2}{6}$. The long-term relative frequency of obtaining this result would approach the theoretical probability of $\frac{2}{6}$ as the number of repetitions grows larger and larger.

This important characteristic of probability experiments is known as the **law of large numbers** which states that as the number of repetitions of an experiment is increased, the relative frequency obtained in the experiment tends to become closer and closer to the theoretical probability. Even though the outcomes do not happen according to any set pattern or order, overall, the long-term observed relative frequency will approach the theoretical probability. (The word **empirical** is often used instead of the word observed.)

It is important to realize that in many situations, the outcomes are not equally likely. A coin or die may be **unfair**, or **biased**. Two math professors in Europe had their statistics students test the Belgian one Euro coin and discovered that in 250 trials, a head was obtained 56% of the time and a tail was obtained 44% of the time. The data seem to show that the coin is not a fair coin; more repetitions would be helpful to draw a more accurate conclusion about such bias. Some dice may be biased. Look at the dice in a game you have at home; the spots on each face are usually small holes carved out and then painted to make the spots visible. Your dice may or may not be biased; it is possible that the outcomes may be affected by the slight weight differences due to the different numbers of holes in the faces. Gambling casinos make a lot of money depending on outcomes from rolling dice, so casino dice are made differently to eliminate bias. Casino dice have flat faces; the holes are completely filled with paint having the same density as the material that the dice are made out of so that each face is equally likely to occur. Later we will learn techniques to use to work with probabilities for events that are not equally likely.

The "OR" Event:

An outcome is in the event A OR B if the outcome is in A or is in B or is in both A and B. For example, let $A = \{1, 2, 3, 4, 5\}$ and $B = \{4, 5, 6, 7, 8\}$. A OR $B = \{1, 2, 3, 4, 5, 6, 7, 8\}$. Notice that 4 and 5 are NOT listed twice.

The "AND" Event:

An outcome is in the event A AND B if the outcome is in both A and B at the same time. For example, let A and B be $\{1, 2, 3, 4, 5\}$ and $\{4, 5, 6, 7, 8\}$, respectively. Then A AND $B = \{4, 5\}$.

The **complement** of event A is denoted A' (read "A prime"). A' consists of all outcomes that are **NOT** in A. Notice that $P(A) + P(A') = 1$. For example, let $S = \{1, 2, 3, 4, 5, 6\}$ and let $A = \{1, 2, 3, 4\}$. Then, $A' = \{5, 6\}$. $P(A) = \frac{4}{6}$, $P(A') = \frac{2}{6}$, and $P(A) + P(A') = \frac{4}{6} + \frac{2}{6} = 1$

The **conditional probability** of A given B is written $P(A|B)$. $P(A|B)$ is the probability that event A will occur given that the event B has already occurred. **A conditional reduces the sample space**. We calculate the probability of A from the reduced sample space B.

The formula to calculate $P(A|B)$ is $P(A|B) = \frac{P(A \text{ AND } B)}{P(B)}$ where $P(B)$ is greater than zero.

For example, suppose we toss one fair, six-sided die. The sample space *S* = {1, 2, 3, 4, 5, 6}. Let *A* = face is 2 or 3 and *B* = face is even (2, 4, 6).
To calculate $P(A|B)$, we count the number of outcomes 2 or 3 in the sample space *B* = {2, 4, 6}. Then we divide that by the number of outcomes *B* (rather than *S*).

We get the same result by using the formula. Remember that *S* has six outcomes.

$$P(A|B) = \frac{P(A \text{ AND } B)}{P(B)} = \frac{\frac{(\text{the number of outcomes that are 2 or 3 and even in } S)}{6}}{\frac{(\text{the number of outcomes that are even in } S)}{6}} = \frac{\frac{1}{6}}{\frac{3}{6}} = \frac{1}{3}$$

Understanding Terminology and Symbols
It is important to read each problem carefully to think about and understand what the events are. Understanding the wording is the first very important step in solving probability problems. Reread the problem several times if necessary. Clearly identify the event of interest. Determine whether there is a condition stated in the wording that would indicate that the probability is conditional; carefully identify the condition, if any.

Example:
Exercise:

Problem:

The sample space *S* is the whole numbers starting at one and less than 20.

a. S = ________________________________

Let event A = the even numbers and event B = numbers greater than 13.

b. A = _______________________, B = _______________________
c. $P(A)$ = ______________, $P(B)$ = _________________
d. A AND B = ______________________, A OR B = _________________
e. $P(A \text{ AND } B)$ = _________, $P(A \text{ OR } B)$ = ______________
f. A' = ______________, $P(A')$ = ______________
g. $P(A) + P(A')$ = _____________
h. $P(A|B)$ = ____________, $P(B|A)$ = ______________; are the probabilities equal?

Solution:

a. $S = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19\}$
b. $A = \{2, 4, 6, 8, 10, 12, 14, 16, 18\}$, $B = \{14, 15, 16, 17, 18, 19\}$
c. $P(A) = \frac{9}{19}$, $P(B) = \frac{6}{19}$
d. A AND $B = \{14,16,18\}$, A OR $B = \{2, 4, 6, 8, 10, 12, 14, 15, 16, 17, 18, 19\}$
e. $P(A \text{ AND } B) = \frac{3}{19}$, $P(A \text{ OR } B) = \frac{12}{19}$
f. $A' = \{1, 3, 5, 7, 9, 11, 13, 15, 17, 19\}$; $P(A') = \frac{10}{19}$
g. $P(A) + P(A') = 1$ $\left(\frac{9}{19} + \frac{10}{19} = 1\right)$
h. $P(A|B) = \frac{P(A \text{ AND } B)}{P(B)} = \frac{3}{6}$, $P(B|A) = \frac{P(A \text{ AND } B)}{P(A)} = \frac{3}{9}$; No

Note:
Try It
Exercise:

Problem:

The sample space S is the ordered pairs of two whole numbers, the first from one to three and the second from one to four (Example: (1, 4)).

a. S = ________________________________

Let event A = the sum is even and event B = the first number is prime.

b. A = _______________________, B = _______________________

c. $P(A)$ = ______________, $P(B)$ = _________________

d. A AND B = ______________________, A OR B = _________________

e. $P(A \text{ AND } B)$ = _________, $P(A \text{ OR } B)$ = ______________

f. B' = ______________, $P(B')$ = ______________

g. $P(A) + P(A')$ = _____________

h. $P(A|B)$ = ____________, $P(B|A)$ = ______________; are the probabilities equal?

Solution:

a. S = {(1,1), (1,2), (1,3), (1,4), (2,1), (2,2), (2,3), (2,4), (3,1), (3,2), (3,3), (3,4)}

b. A = {(1,1), (1,3), (2,2), (2,4), (3,1), (3,3)}

B = {(2,1), (2,2), (2,3), (2,4), (3,1), (3,2), (3,3), (3,4)}

c. $P(A) = \frac{1}{2}$, $P(B) = \frac{2}{3}$

d. A AND B = {(2,2), (2,4), (3,1), (3,3)}

A OR B = {(1,1), (1,3), (2,1), (2,2), (2,3), (2,4), (3,1), (3,2), (3,3), (3,4)}

e. $P(A \text{ AND } B) = \frac{1}{3}$, $P(A \text{ OR } B) = \frac{5}{6}$

f. B' = {(1,1), (1,2), (1,3), (1,4)}, $P(B') = \frac{1}{3}$

g. $P(B) + P(B') = 1$
h. $P(A|B) = \frac{P(A \text{ AND } B)}{P(B)} = \frac{1}{2}$, $P(B|A) = \frac{P(A \text{ AND } B)}{P(B)} = \frac{2}{3}$; No.

Example:
Exercise:

Problem:

A fair, six-sided die is rolled. Describe the sample space *S*, identify each of the following events with a subset of *S* and compute its probability (an outcome is the number of dots that show up).

a. Event T = the outcome is two.
b. Event A = the outcome is an even number.
c. Event B = the outcome is less than four.
d. The complement of A.
e. A GIVEN B
f. B GIVEN A
g. A AND B
h. A OR B
i. A OR B'
j. Event N = the outcome is a prime number.
k. Event I = the outcome is seven.

Solution:

a. $T = \{2\}$, $P(T) = \frac{1}{6}$
b. $A = \{2, 4, 6\}$, $P(A) = \frac{1}{2}$
c. $B = \{1, 2, 3\}$, $P(B) = \frac{1}{2}$
d. $A' = \{1, 3, 5\}$, $P(A') = \frac{1}{2}$
e. $A|B = \{2\}$, $P(A|B) = \frac{1}{3}$
f. $B|A = \{2\}$, $P(B|A) = \frac{1}{3}$

g. A AND $B = \{2\}$, $P(A \text{ AND } B) = \frac{1}{6}$
h. A OR $B = \{1, 2, 3, 4, 6\}$, $P(A \text{ OR } B) = \frac{5}{6}$
i. A OR $B' = \{2, 4, 5, 6\}$, $P(A \text{ OR } B') = \frac{2}{3}$
j. $N = \{2, 3, 5\}$, $P(N) = \frac{1}{2}$
k. A six-sided die does not have seven dots. $P(7) = 0$.

Example:

The table below describes the distribution of a random sample S of 100 individuals, organized by gender and whether they are right- or left-handed.

	Right-handed	**Left-handed**
Males	43	9
Females	44	4

Exercise:

Problem:

Let's denote the events M = the subject is male, F = the subject is female, R = the subject is right-handed, L = the subject is left-handed. Compute the following probabilities:

a. $P(M)$
b. $P(F)$
c. $P(R)$

d. $P(L)$
e. $P(M \text{ AND } R)$
f. $P(F \text{ AND } L)$
g. $P(M \text{ OR } F)$
h. $P(M \text{ OR } R)$
i. $P(F \text{ OR } L)$
j. $P(M')$
k. $P(R|M)$
l. $P(F|L)$
m. $P(L|F)$

Solution:

a. $P(M) = 0.52$
b. $P(F) = 0.48$
c. $P(R) = 0.87$
d. $P(L) = 0.13$
e. $P(M \text{ AND } R) = 0.43$
f. $P(F \text{ AND } L) = 0.04$
g. $P(M \text{ OR } F) = 1$
h. $P(M \text{ OR } R) = 0.96$
i. $P(F \text{ OR } L) = 0.57$
j. $P(M') = 0.48$
k. $P(R|M) = 0.8269$ (rounded to four decimal places)
l. $P(F|L) = 0.3077$ (rounded to four decimal places)
m. $P(L|F) = 0.0833$

References

“Countries List by Continent.” Worldatlas, 2013. Available online at http://www.worldatlas.com/cntycont.htm (accessed May 2, 2013).

Section Review

In this module we learned the basic terminology of probability. The set of all possible outcomes of an experiment is called the **sample space**. **Events** are subsets of the sample space, and they are assigned a probability that is a number between zero and one, inclusive.

An outcome is in the event *A* OR *B* if the outcome is in *A* or is in *B* or is in both *A* and *B*.

An outcome is in the event *A* AND *B* if the outcome is in both *A* and *B* at the same time.

The **complement** of event *A*, denoted *A′*, consists of all outcomes that are **NOT** in *A*.

The **conditional probability** of *A* given *B* is the probability that event *A* will occur given that the event *B* has already occurred.

Formula Review

A and *B* are events

$P(S) = 1$ where *S* is the sample space

$0 \leq P(A) \leq 1$

$P\left(A\middle|B\right) = \frac{P(A \text{ AND } B)}{P(B)}$

$P(A) + P(A') = 1$

Exercise:

Problem:

In a particular college class, there are male and female students. Some students have long hair and some students have short hair. Write the **symbols** for the probabilities of the events for parts a through j. (Note that you cannot find numerical answers here. You were not given enough information to find any probability values yet; concentrate on understanding the symbols.)

- Let F be the event that a student is female.
- Let M be the event that a student is male.
- Let S be the event that a student has short hair.
- Let L be the event that a student has long hair.

a. The probability that a student does not have long hair.
b. The probability that a student is male or has short hair.
c. The probability that a student is a female and has long hair.
d. The probability that a student is male, given that the student has long hair.
e. The probability that a student has long hair, given that the student is male.
f. Of all the female students, the probability that a student has short hair.
g. Of all students with long hair, the probability that a student is female.
h. The probability that a student is female or has long hair.
i. The probability that a randomly selected student is a male student with short hair.
j. The probability that a student is female.

Solution:

a. $P(L') = P(S)$
b. $P(M \text{ OR } S)$
c. $P(F \text{ AND } L)$
d. $P(M|L)$

e. $P(L|M)$
f. $P(S|F)$
g. $P(F|L)$
h. $P(F \text{ OR } L)$
i. $P(M \text{ AND } S)$
j. $P(F)$

Use the following information to answer the next four exercises. A box is filled with several party favors. It contains 12 hats, 15 noisemakers, ten finger traps, and five bags of confetti.
Let H = the event of getting a hat.
Let N = the event of getting a noisemaker.
Let F = the event of getting a finger trap.
Let C = the event of getting a bag of confetti.

Exercise:

Problem: Find $P(H)$.

Exercise:

Problem: Find $P(N)$.

Solution:

$P(N) = \frac{15}{42} = \frac{5}{14} = 0.36$

Exercise:

Problem: Find $P(F)$.

Exercise:

Problem: Find $P(C)$.

Solution:

$P(C) = \frac{5}{42} = 0.12$

Use the following information to answer the next six exercises. A jar of 150 jelly beans contains 22 red jelly beans, 38 yellow, 20 green, 28 purple, 26 blue, and the rest are orange.
Let B = the event of getting a blue jelly bean
Let G = the event of getting a green jelly bean.
Let O = the event of getting an orange jelly bean.
Let P = the event of getting a purple jelly bean.
Let R = the event of getting a red jelly bean.
Let Y = the event of getting a yellow jelly bean.

Exercise:

Problem: Find $P(B)$.

Exercise:

Problem: Find $P(G)$.

Solution:

$P(G) = \frac{20}{150} = \frac{2}{15} = 0.13$

Exercise:

Problem: Find $P(P)$.

Exercise:

Problem: Find $P(R)$.

Solution:

$P(R) = \frac{22}{150} = \frac{11}{75} = 0.15$

Exercise:

Problem: Find $P(Y)$.

Exercise:

Problem: Find $P(O)$.

Solution:

$P(O)$ = $\frac{150-22-38-20-28-26}{150}$ = $\frac{16}{150}$ = $\frac{8}{75}$ = 0.11

Use the following information to answer the next six exercises. There are 23 countries in North America, 12 countries in South America, 47 countries in Europe, 44 countries in Asia, 54 countries in Africa, and 14 in Oceania (Pacific Ocean region).
Let *A* = the event that a country is in Asia.
Let *E* = the event that a country is in Europe.
Let *F* = the event that a country is in Africa.
Let *N* = the event that a country is in North America.
Let *O* = the event that a country is in Oceania.
Let *S* = the event that a country is in South America.

Exercise:

Problem: Find $P(A)$.

Exercise:

Problem: Find $P(E)$.

Solution:

$P(E)$ = $\frac{47}{194}$ = 0.24

Exercise:

Problem: Find $P(F)$.

Exercise:

Problem: Find $P(N)$.

Solution:

$P(N)) = \frac{23}{194} = 0.12$

Exercise:

Problem: Find $P(O)$.

Exercise:

Problem: Find $P(S)$.

Solution:

$P(S) = \frac{12}{194} = \frac{6}{97} = 0.06$

Exercise:

Problem:

What is the probability of drawing a red card in a standard deck of 52 cards?

Exercise:

Problem:

What is the probability of drawing a club in a standard deck of 52 cards?

Solution:

$\frac{13}{52} = \frac{1}{4} = 0.25$

Exercise:

Problem:

What is the probability of rolling an even number of dots with a fair, six-sided die numbered one through six?

Exercise:

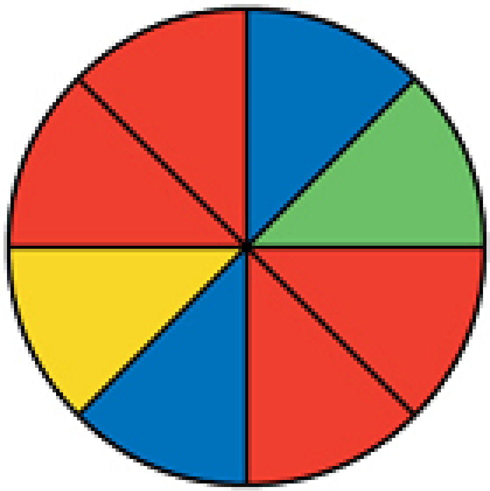
Problem:

What is the probability of rolling a prime number of dots with a fair, six-sided die numbered one through six?

Solution:

$\frac{3}{6} = \frac{1}{2} = 0.5$

Use the following information to answer the next two exercises. You see a game at a local fair. You have to throw a dart at a color wheel. Each section on the color wheel is equal in area.



Let B = the event of landing on blue.
Let R = the event of landing on red.

Let *G* = the event of landing on green.
Let *Y* = the event of landing on yellow.
Exercise:

Problem: If you land on *Y*, you get the biggest prize. Find $P(Y)$.

Exercise:

Problem: If you land on red, you don't get a prize. What is $P(R)$?

Solution:

$P(R) = \frac{4}{8} = 0.5$

Use the following information to answer the next ten exercises. On a baseball team, there are infielders and outfielders. Some players are great hitters, and some players are not great hitters.
Let *I* = the event that a player in an infielder.
Let *O* = the event that a player is an outfielder.
Let *H* = the event that a player is a great hitter.
Let *N* = the event that a player is not a great hitter.
Exercise:

Problem:

Write the symbols for the probability that a player is not an outfielder.

Exercise:

Problem:

Write the symbols for the probability that a player is an outfielder or is a great hitter.

Solution:

$P(O \text{ OR } H)$

Exercise:

Problem:

Write the symbols for the probability that a player is an infielder and is not a great hitter.

Exercise:

Problem:

Write the symbols for the probability that a player is a great hitter, given that the player is an infielder.

Solution:

$P(H|I)$

Exercise:

Problem:

Write the symbols for the probability that a player is an infielder, given that the player is a great hitter.

Exercise:

Problem:

Write the symbols for the probability that of all the outfielders, a player is not a great hitter.

Solution:

$P(N|O)$

Exercise:

Problem:

Write the symbols for the probability that of all the great hitters, a player is an outfielder.

Exercise:

Problem:

Write the symbols for the probability that a player is an infielder or is not a great hitter.

Solution:

$P(I \text{ OR } N)$

Exercise:

Problem:

Write the symbols for the probability that a player is an outfielder and is a great hitter.

Exercise:

Problem:

Write the symbols for the probability that a player is an infielder.

Solution:

$P(I)$

Exercise:

Problem: What is the word for the set of all possible outcomes?

Exercise:

Problem: What is conditional probability?

Solution:

The likelihood that an event will occur given that another event has already occurred.

Exercise:

Problem:

A shelf holds 12 books. Eight are fiction and the rest are nonfiction. Each is a different book with a unique title. The fiction books are numbered one to eight. The nonfiction books are numbered one to four. Randomly select one book
Let F = event that book is fiction
Let N = event that book is nonfiction
What is the sample space?

Exercise:

Problem:

What is the sum of the probabilities of an event and its complement?

Solution:

1

Use the following information to answer the next two exercises. You are rolling a fair, six-sided number cube. Let E = the event that it lands on an even number. Let M = the event that it lands on a multiple of three.

Exercise:

Problem: What does $P(E|M)$ mean in words?

Exercise:

Problem: What does $P(E \text{ OR } M)$ mean in words?

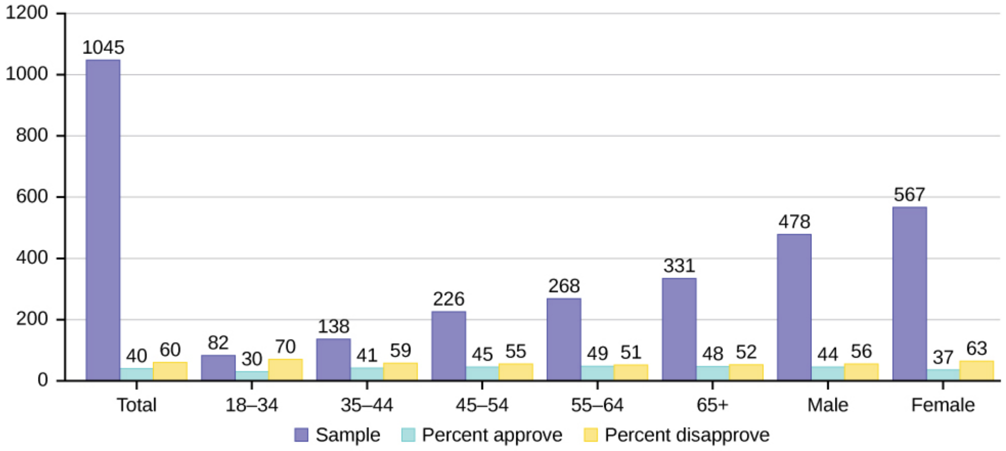
Solution:

the probability of landing on an even number or a multiple of three

Homework

Exercise:

Problem:



The graph in the figure above displays the sample sizes and percentages of people in different age and gender groups who were polled concerning their approval of Mayor Ford's actions in office. The total number in the sample of all the age groups is 1,045.

a. Define three events in the graph.
b. Describe in words what the entry 40 means.
c. Describe in words the complement of the entry in question 2.
d. Describe in words what the entry 30 means.
e. Out of the males and females, what percent are males?
f. Out of the females, what percent disapprove of Mayor Ford?
g. Out of all the age groups, what percent approve of Mayor Ford?
h. Find $P(\text{Approve}|\text{Male})$.
i. Out of the age groups, what percent are more than 44 years old?
j. Find $P(\text{Approve}|\text{Age} < 35)$.

Exercise:

Problem:

Explain what is wrong with the following statements. Use complete sentences.

a. If there is a 60% chance of rain on Saturday and a 70% chance of rain on Sunday, then there is a 130% chance of rain over the weekend.
b. The probability that a baseball player hits a home run is greater than the probability that he gets a successful hit.

Solution:

a. You can't calculate the joint probability knowing the probability of both events occurring, which is not in the information given; the probabilities should be multiplied, not added; and probability is never greater than 100%
b. A home run by definition is a successful hit, so he has to have at least as many successful hits as home runs.

Glossary

Conditional Probability
: the likelihood that an event will occur given that another event has already occurred

Equally Likely
: Each outcome of an experiment has the same probability.

Event
: a subset of the set of all outcomes of an experiment; the set of all outcomes of an experiment is called a **sample space** and is usually denoted by *S*. An event is an arbitrary subset in *S*. It can contain one outcome, two outcomes, no outcomes (empty subset), the entire sample space, and the like. Standard notations for events are capital letters such as *A*, *B*, *C*, and so on.

Experiment
: a planned activity carried out under controlled conditions

Outcome

a particular result of an experiment

Probability

a number between zero and one, inclusive, that gives the likelihood that a specific event will occur. Let S denote the sample space and A is an event in S. Then:

- $0 \leq P(A) \leq 1$
- $P(S) = 1$

Sample Space

the set of all possible outcomes of an experiment

The AND Event

An outcome is in the event A AND B if the outcome is in both A AND B at the same time.

The Complement Event

The complement of event A consists of all outcomes that are NOT in A.

The Conditional Probability of A GIVEN B

$P(A|B)$ is the probability that event A will occur given that the event B has already occurred.

The Or Event

An outcome is in the event A OR B if the outcome is in A or is in B or is in both A and B.

Independent and Mutually Exclusive Events

Independent and mutually exclusive do **not** mean the same thing.

Independent Events

Two events *A* and *B* are **independent** if the knowledge that one occurred does not affect the chance the other occurs. For example, the outcomes of two rolls of a fair die are independent events. The outcome of the first roll does not change the probability for the outcome of the second roll. If two events are NOT independent, then we say that they are **dependent**.

Two events are independent if the following are true:

- $P(A|B) = P(A)$
- $P(B|A) = P(B)$
- $P(A \text{ AND } B) = P(A)P(B)$

To show two events are independent, you must show **only one** of the above conditions.

Sampling may be done **with replacement** or **without replacement**.

- **With replacement**: If each member of a population is replaced after it is picked, then that member has the possibility of being chosen more than once. When sampling is done with replacement, then events are considered to be independent, meaning the result of the first pick will not change the probabilities for the second pick.
- **Without replacement**: When sampling is done without replacement, each member of a population may be chosen only once. In this case, the probabilities for the second pick are affected by the result of the first pick. The events are considered to be dependent or not independent.

If it is not known whether *A* and *B* are independent or dependent, **assume they are dependent until you can show otherwise**.

Example:
You have a fair, well-shuffled deck of 52 cards. It consists of four suits. The suits are clubs, diamonds, hearts and spades. There are 13 cards in each suit consisting of *A* (ace), 2, 3, 4, 5, 6, 7, 8, 9, 10, *J* (jack), *Q* (queen), *K* (king) of that suit.

a. Sampling with replacement:

Suppose you pick three cards with replacement. The first card you pick out of the 52 cards is the *Q* of spades. You put this card back, reshuffle the cards and pick a second card from the 52-card deck. It is the ten of clubs. You put this card back, reshuffle the cards and pick a third card from the 52-card deck. This time, the card is the *Q* of spades again. Your picks are {*Q* of spades, ten of clubs, *Q* of spades}. You have picked the *Q* of spades twice. You pick each card from the 52-card deck.

b. Sampling without replacement:

Suppose you pick three cards without replacement. The first card you pick out of the 52 cards is the *K* of hearts. You put this card aside and pick the second card from the 51 cards remaining in the deck. It is the three of diamonds. You put this card aside and pick the third card from the remaining 50 cards in the deck. The third card is the *J* of spades. Your picks are {*K* of hearts, three of diamonds, *J* of spades}. Because you have picked the cards without replacement, you cannot pick the same card twice.

Note:
Try It

Exercise:

Problem:

You have a fair, well-shuffled deck of 52 cards. It consists of four suits. The suits are clubs, diamonds, hearts and spades. There are 13 cards in each suit consisting of *A* (ace), 2, 3, 4, 5, 6, 7, 8, 9, 10, *J* (jack), *Q* (queen), *K* (king) of that suit. Three cards are picked at random.

a. Suppose you know that the picked cards are *Q* of spades, *K* of hearts and *Q* of spades. Can you decide if the sampling was with or without replacement?
b. Suppose you know that the picked cards are *Q* of spades, *K* of hearts, and *J* of spades. Can you decide if the sampling was with or without replacement?

Solution:

a. With replacement
b. No

Example:
Exercise:

Problem:

You have a fair, well-shuffled deck of 52 cards. It consists of four suits. The suits are clubs, diamonds, hearts, and spades. There are 13 cards in each suit consisting of *A* (ace), 2, 3, 4, 5, 6, 7, 8, 9, 10, *J* (jack), *Q* (queen), and *K* (king) of that suit. *S* = spades, *H* = Hearts, *D* = Diamonds, *C* = Clubs.

a. Suppose you pick four cards, but do not put any cards back into the deck. Your cards are *QS*, *AD*, *AC*, *QD*.

b. Suppose you pick four cards and put each card back before you pick the next card. Your cards are *KH*, 7*D*, 6*D*, *KH*.

Which of a. or b. did you sample with replacement and which did you sample without replacement?

Solution:

a. Without replacement; b. With replacement

Note:
Try It
Exercise:

Problem:

You have a fair, well-shuffled deck of 52 cards. It consists of four suits. The suits are clubs, diamonds, hearts, and spades. There are 13 cards in each suit consisting of *A* (ace), 2, 3, 4, 5, 6, 7, 8, 9, 10, *J* (jack), *Q* (queen), and *K* (king) of that suit. *S* = spades, *H* = Hearts, *D* = Diamonds, *C* = Clubs. Suppose that you sample four cards without replacement. Which of the following outcomes are possible? Answer the same question for sampling with replacement.

a. *QS*, *AD*, *AC*, *QD*
b. *KH*, 7*D*, 6*D*, *KH*
c. *QS*, 7*D*, 6*D*, *KS*

Solution:

without replacement: a. Possible; b. Impossible, c. Possible

with replacement: a. Possible; b. Possible, c. Possible

Mutually Exclusive Events

A and B are **mutually exclusive** events if they cannot occur at the same time. This means that A and B do not share any outcomes and $P(A \text{ AND } B) = 0$.

For example, suppose the sample space S = {1, 2, 3, 4, 5, 6, 7, 8, 9, 10}. Let A = {1, 2, 3, 4, 5}, B = {4, 5, 6, 7, 8}, and C = {7, 9}. A AND B = {4, 5}. $P(A \text{ AND } B) = \frac{2}{10}$ and is not equal to zero. Therefore, A and B are not mutually exclusive. A and C do not have any numbers in common so $P(A \text{ AND } C) = 0$. Therefore, A and C are mutually exclusive.

If it is not known whether A and B are mutually exclusive, **assume they are not until you can show otherwise**. The following examples illustrate these definitions and terms.

Example:
Flip two fair coins. (This is an experiment.)
The sample space is {HH, HT, TH, TT} where T = tails and H = heads. The outcomes are HH, HT, TH, and TT. The outcomes HT and TH are different. The HT means that the first coin showed heads and the second coin showed tails. The TH means that the first coin showed tails and the second coin showed heads.

- Let A = the event of getting **at most one tail**. (At most one tail means zero or one tail.) Then A can be written as {HH, HT, TH}. The outcome HH shows zero tails. HT and TH each show one tail.
- Let B = the event of getting all tails. B can be written as {TT}. B is the **complement** of A, so $B = A'$. Also, $P(A) + P(B) = P(A) + P(A') = 1$.
- The probabilities for A and for B are $P(A) = \frac{3}{4}$ and $P(B) = \frac{1}{4}$.
- Let C = the event of getting all heads. C = {HH}. Since B = {TT}, $P(B \text{ AND } C) = 0$. B and C are mutually exclusive. (B and C have no

members in common because you cannot have all tails and all heads at the same time.)
- Let D = event of getting **more than one** tail. D = {TT}. $P(D) = \frac{1}{4}$
- Let E = event of getting a head on the first roll. (This implies you can get either a head or tail on the second roll.) E = {HT, HH}. $P(E) = \frac{2}{4}$
- Find the probability of getting **at least one** (one or two) tail in two flips. Let F = event of getting at least one tail in two flips. F = {HT, TH, TT}. $P(F) = \frac{3}{4}$

Note:
Try It
Exercise:

Problem:

Draw two cards from a standard 52-card deck with replacement. Find the probability of getting at least one black card.

Solution:
Try It Solutions

The sample space of drawing two cards with replacement from a standard 52-card deck with respect to color is {BB, BR, RB, RR}.

Event A = Getting at least one black card = {BB, BR, RB}

$P(A) = \frac{3}{4} = 0.75$

Example:
Exercise:

Problem: Flip two fair coins. Find the probabilities of the events.

a. Let *F* = the event of getting at most one tail (zero or one tail).
b. Let *G* = the event of getting two faces that are the same.
c. Let *H* = the event of getting a head on the first flip followed by a head or tail on the second flip.
d. Are *F* and *G* mutually exclusive?
e. Let *J* = the event of getting all tails. Are *J* and *H* mutually exclusive?

Solution:

Look at the sample space in [link].

a. Zero (0) or one (1) tails occur when the outcomes *HH*, *TH*, *HT* show up. $P(F) = \frac{3}{4}$
b. Two faces are the same if *HH* or *TT* show up. $P(G) = \frac{2}{4}$
c. A head on the first flip followed by a head or tail on the second flip occurs when *HH* or *HT* show up. $P(H) = \frac{2}{4}$
d. *F* and *G* share *HH* so $P(F \text{ AND } G)$ is not equal to zero (0). *F* and *G* are not mutually exclusive.
e. Getting all tails occurs when tails shows up on both coins (*TT*). *H*'s outcomes are *HH* and *HT*.

J and *H* have nothing in common so $P(J \text{ AND } H) = 0$. *J* and *H* are mutually exclusive.

Note:
Try It
Exercise:

Problem:

A box has two balls, one white and one red. We select one ball, put it back in the box, and select a second ball (sampling with replacement). Find the probability of the following events:

a. Let F = the event of getting the white ball twice.
b. Let G = the event of getting two balls of different colors.
c. Let H = the event of getting white on the first pick.
d. Are F and G mutually exclusive?
e. Are G and H mutually exclusive?

Solution:

a. $P(F) = \frac{1}{4}$
b. $P(G) = \frac{1}{2}$
c. $P(H) = \frac{1}{2}$
d. Yes
e. No

Example:

Roll one fair, six-sided die. The sample space is {1, 2, 3, 4, 5, 6}. Let event A = a face is odd. Then A = {1, 3, 5}. Let event B = a face is even. Then B = {2, 4, 6}.

- Find the complement of A, A'. The complement of A, A', is B because A and B together make up the sample space. $P(A) + P(B) = P(A) + P(A') = 1$. Also, $P(A) = \frac{3}{6}$ and $P(B) = \frac{3}{6}$.
- Let event C = odd faces larger than two. Then C = {3, 5}. Let event D = all even faces smaller than five. Then D = {2, 4}. $P(C \text{ AND } D) = 0$ because you cannot have an odd and even face at the same time. Therefore, C and D are mutually exclusive events.
- Let event E = all faces less than five. E = {1, 2, 3, 4}.

Exercise:

Problem:

Are C and E mutually exclusive events? (Answer yes or no.) Why or why not?

Solution:

No. C = {3, 5} and E = {1, 2, 3, 4}. $P(C \text{ AND } E) = \frac{1}{6}$. To be mutually exclusive, $P(C \text{ AND } E)$ must be zero.

- Find $P(C|A)$. This is a conditional probability. Recall that the event C is {3, 5} and event A is {1, 3, 5}. To find $P(C|A)$, find the probability of C using the sample space A. You have reduced the sample space from the original sample space {1, 2, 3, 4, 5, 6} to {1, 3, 5}. So, $P(C|A) = \frac{2}{3}$.

Note:
Try It
Exercise:

Problem:

Let event A = learning Spanish. Let event B = learning German. Then A AND B = learning Spanish and German. Suppose $P(A) = 0.4$ and $P(B) = 0.2$. $P(A \text{ AND } B) = 0.08$. Are events A and B independent? Hint: You must show ONE of the following:

- $P(A|B) = P(A)$
- $P(B|A) = P(B)$
- $P(A \text{ AND } B) = P(A)P(B)$

Solution:

$$P(A|B) = \frac{P(A \text{ AND } B)}{P(B)} = \frac{0.08}{0.2} = 0.4 = P(A)$$

The events are independent because $P(A|B) = P(A)$.

Example:
Let event G = taking a math class. Let event H = taking a science class. Then, G AND H = taking a math class and a science class. Suppose $P(G) = 0.6$, $P(H) = 0.5$, and $P(G \text{ AND } H) = 0.3$. Are G and H independent?
If G and H are independent, then you must show **ONE** of the following:

- $P(G|H) = P(G)$
- $P(H|G) = P(H)$
- $P(G \text{ AND } H) = P(G)P(H)$

Note:
NOTE
The choice you make depends on the information you have. You could choose any of the methods here because you have the necessary information.

Exercise:

Problem: a. Show that $P(G|H) = P(G)$.

Solution:

$$P(G|H) = \frac{P(G \text{ AND } H)}{P(H)} = \frac{0.3}{0.5} = 0.6 = P(G)$$

Exercise:

Problem: b. Show $P(G \text{ AND } H) = P(G)P(H)$.

Solution:

$P(G)P(H)$ = (0.6)(0.5) = 0.3 = $P(G \text{ AND } H)$.

Since G and H are independent, knowing that a person is taking a science class does not change the chance that he or she is taking a math class. If the two events had not been independent (that is, they are dependent) then knowing that a person is taking a science class would change the chance he or she is taking math. For practice, show that $P(H|G) = P(H)$ to show that G and H are independent events.

Note:
Try It
Exercise:

Problem:

In a bag, there are six red marbles and four green marbles. The red marbles are marked with the numbers 1, 2, 3, 4, 5, and 6. The green marbles are marked with the numbers 1, 2, 3, and 4.

- R = a red marble
- G = a green marble
- O = an odd-numbered marble
- The sample space is $S = \{R1, R2, R3, R4, R5, R6, G1, G2, G3, G4\}$.

S has ten outcomes. What is $P(G \text{ AND } O)$?

Solution:

Event G and $O = \{G1, G3\}$

$P(G \text{ AND } H) = \frac{2}{10} = 0.2$

Example:
Exercise:

Problem:

Let event *C* = taking an English class. Let event *D* = taking a speech class.

Suppose $P(C) = 0.75$, $P(D) = 0.3$, $P(C|D) = 0.75$ and $P(C \text{ AND } D) = 0.225$.

Justify your answers to the following questions numerically.

a. Are *C* and *D* independent?
b. Are *C* and *D* mutually exclusive?
c. What is $P(D|C)$?

Solution:

a. Yes, because $P(C|D) = P(C)$.
b. No, because $P(C \text{ AND } D)$ is not equal to zero.
c. $P(D|C) = \frac{P(C \text{ AND } D)}{P(C)} = \frac{0.225}{0.75} = 0.3$

Note:
Try It
Exercise:

Problem:

A student goes to the library. Let events B = the student checks out a book and D = the student checks out a DVD. Suppose that $P(B)$ = 0.40, $P(D)$ = 0.30 and $P(B \text{ AND } D)$ = 0.20.

a. Find $P(B|D)$.
b. Find $P(D|B)$.
c. Are B and D independent?
d. Are B and D mutually exclusive?

Solution:

a. $P(B|D)$ = 0.6667
b. $P(D|B)$ = 0.5
c. No
d. No

Example:

In a box there are three red cards and five blue cards. The red cards are marked with the numbers 1, 2, and 3, and the blue cards are marked with the numbers 1, 2, 3, 4, and 5. The cards are well-shuffled. You reach into the box (you cannot see into it) and draw one card.
Let R = red card is drawn, B = blue card is drawn, E = even-numbered card is drawn.
The sample space S = $R1$, $R2$, $R3$, $B1$, $B2$, $B3$, $B4$, $B5$. S has eight outcomes.

- $P(R) = \frac{3}{8}$. $P(B) = \frac{5}{8}$. $P(R \text{ AND } B) = 0$. (You cannot draw one card that is both red and blue.)
- $P(E) = \frac{3}{8}$. (There are three even-numbered cards, $R2$, $B2$, and $B4$.)

- $P(E|B) = \frac{2}{5}$. (There are five blue cards: $B1$, $B2$, $B3$, $B4$, and $B5$. Out of the blue cards, there are two even cards; $B2$ and $B4$.)
- $P(B|E) = \frac{2}{3}$. (There are three even-numbered cards: $R2$, $B2$, and $B4$. Out of the even-numbered cards, two are blue; $B2$ and $B4$.)
- The events R and B are mutually exclusive because $P(R \text{ AND } B) = 0$.
- Let G = card with a number greater than 3. $G = \{B4, B5\}$. $P(G) = \frac{2}{8}$. Let H = blue card numbered between one and four, inclusive. $H = \{B1, B2, B3, B4\}$. $P(G|H) = \frac{1}{4}$. (The only card in H that has a number greater than three is $B4$.) Since $\frac{2}{8} = \frac{1}{4}$, $P(G) = P(G|H)$, which means that G and H are independent.

Note:
Try It
Exercise:

Problem: In a basketball arena,

- 70% of the fans are rooting for the home team.
- 25% of the fans are wearing blue.
- 20% of the fans are wearing blue and are rooting for the away team.
- Of the fans rooting for the away team, 67% are wearing blue.

Let A be the event that a fan is rooting for the away team.
Let B be the event that a fan is wearing blue.
Are the events of rooting for the away team and wearing blue independent? Are they mutually exclusive?

Solution:

$P(B|A) = 0.67$

$P(B) = 0.25$

So $P(B)$ does not equal $P(B|A)$ which means that B and A are not independent (wearing blue and rooting for the away team are not independent). They are also not mutually exclusive, because $P(B \text{ AND } A) = 0.20$, not 0.

Example:

In a particular college class, 60% of the students are female. Fifty percent of all students in the class have long hair. Forty-five percent of the students are female and have long hair. Of the female students, 75% have long hair. Let F be the event that a student is female. Let L be the event that a student has long hair. One student is picked randomly. Are the events of being female and having long hair independent?

- The following probabilities are given in this example:
- $P(F) = 0.60$; $P(L) = 0.50$
- $P(F \text{ AND } L) = 0.45$
- $P(L|F) = 0.75$

Note:

NOTE

The choice you make depends on the information you have. You could use the first or last condition on the list for this example. You do not know $P(F|L)$ yet, so you cannot use the second condition.

Solution 1

Check whether $P(F \text{ AND } L) = P(F)P(L)$. We are given that $P(F \text{ AND } L) = 0.45$, but $P(F)P(L) = (0.60)(0.50) = 0.30$. The events of being female and having long hair are not independent because $P(F \text{ AND } L)$ does not equal $P(F)P(L)$.

Solution 2

Check whether $P(L\,|F)$ equals $P(L)$. We are given that $P(L\,|F)$ = 0.75, but $P(L)$ = 0.50; they are not equal. The events of being female and having long hair are not independent.

Interpretation of Results

The events of being female and having long hair are not independent; knowing that a student is female changes the probability that a student has long hair.

Note:

Try It

Exercise:

Problem:

Mark is deciding which route to take to work. His choices are *I* = the Interstate and *F* = Fifth Street.

- $P(I)$ = 0.44 and $P(F)$ = 0.56
- $P(I \text{ AND } F)$ = 0 because Mark will take only one route to work.

Are events *I* and *F* independent events?

Solution:

No. If *I* and *F* are independent events, then $P(I \text{ AND } F)$ must equal $P(I)P(F)$.

$P(I \text{ AND } F)$ = 0. $P(I)P(F)$ = (0.44)(0.56) = 0.2464.
$P(I \text{ AND } F)$ does not equal $P(I)P(F)$.

Example:

Exercise:

Problem:

a. Toss one fair coin (the coin has two sides, *H* and *T*). The outcomes are ________. Count the outcomes. There are ____ outcomes.
b. Toss one fair, six-sided die (the die has 1, 2, 3, 4, 5 or 6 dots on a side). The outcomes are ________________. Count the outcomes. There are ___ outcomes.
c. Multiply the two numbers of outcomes. The answer is _______.
d. If you flip one fair coin and follow it with the toss of one fair, six-sided die, the answer to c is the number of outcomes (size of the sample space). What are the outcomes? (Hint: Two of the outcomes are *H*1 and *T*6.)
e. Event *A* = heads (*H*) on the coin followed by an even number (2, 4, 6) on the die.
A = {_________________}. Find $P(A)$.
f. Event *B* = heads on the coin followed by a three on the die. *B* = {________}. Find $P(B)$.
g. Are *A* and *B* mutually exclusive? (Hint: What is $P(A \text{ AND } B)$? If $P(A \text{ AND } B) = 0$, then *A* and *B* are mutually exclusive.)
h. Are *A* and *B* independent? (Hint: Does $P(A \text{ AND } B) = P(A)P(B)$? If $P(A \text{ AND } B) = P(A)P(B)$, then *A* and *B* are independent. If not, then they are dependent).

Solution:

a. *H* and *T*; 2
b. 1, 2, 3, 4, 5, 6; 6
c. 2(6) = 12
d. *T*1, *T*2, *T*3, *T*4, *T*5, *T*6, *H*1, *H*2, *H*3, *H*4, *H*5, *H*6
e. *A* = {*H*2, *H*4, *H*6}; $P(A) = \frac{3}{12}$
f. *B* = {*H*3}; $P(B) = \frac{1}{12}$
g. Yes, because $P(A \text{ AND } B) = 0$
h. $P(A \text{ AND } B) = 0$. $P(A)P(B) = \left(\frac{3}{12}\right)\left(\frac{1}{12}\right)$. $P(A \text{ AND } B)$ does not equal $P(A)P(B)$, so *A* and *B* are dependent.

Note:
Try It
Exercise:

Problem:

A box has two balls, one white and one red. We select one ball, put it back in the box, and select a second ball (sampling with replacement). Let *T* be the event of getting the white ball twice, *F* the event of picking the white ball first, *S* the event of picking the white ball in the second drawing.

a. Compute $P(T)$.
b. Compute $P(T|F)$.
c. Are *T* and *F* independent?.
d. Are *F* and *S* mutually exclusive?
e. Are *F* and *S* independent?

Solution:

a. $P(T) = \frac{1}{4}$
b. $P(T|F) = \frac{1}{2}$
c. No
d. No
e. Yes

References

Lopez, Shane, Preety Sidhu. “U.S. Teachers Love Their Lives, but Struggle in the Workplace.” Gallup Wellbeing, 2013. http://www.gallup.com/poll/161516/teachers-love-lives-struggle-workplace.aspx (accessed May 2, 2013).

Section Review

Two events *A* and *B* are independent if the knowledge that one occurred does not affect the chance the other occurs. If two events are not independent, then we say that they are dependent.

In sampling with replacement, each member of a population is replaced after it is picked, so that member has the possibility of being chosen more than once, and the events are considered to be independent. In sampling without replacement, each member of a population may be chosen only once, and the events are considered not to be independent. When events do not share outcomes, they are mutually exclusive of each other.

Two events are mutually exclusive if they cannot occur at the same time.

Formula Review

If *A* and *B* are independent, then $P(A \text{ AND } B)$ = $P(A)P(B)$, $P(A|B)$ = $P(A)$, and $P(B|A)$ = $P(B)$.

If *A* and *B* are mutually exclusive, then $P(A \text{ AND } B) = 0$.

Exercise:

Problem:

E and *F* are mutually exclusive events. $P(E)$ = 0.4; $P(F)$ = 0.5. Find $P(E|F)$.

Exercise:

Problem: *J* and *K* are independent events. $P(J|K)$ = 0.3. Find $P(J)$.

Solution:

$P(J)$ = 0.3

Exercise:

Problem:

U and V are mutually exclusive events. $P(U) = 0.26$; $P(V) = 0.37$. Find:

a. $P(U \text{ AND } V) =$
b. $P(U|V) =$

Exercise:

Problem:

Q and R are independent events. $P(Q) = 0.4$ and $P(Q \text{ AND } R) = 0.1$. Find $P(R)$.

Solution:

$P(Q \text{ AND } R) = P(Q)P(R)$

$0.1 = (0.4)P(R)$

$P(R) = 0.25$

Homework

Exercise:

Problem: Consider the following scenario:
Let $P(C) = 0.4$.
Let $P(D) = 0.5$.
Let $P(C|D) = 0.6$.

a. Find $P(C \text{ AND } D)$.
b. Are C and D mutually exclusive? Why or why not?
c. Are C and D independent events? Why or why not?
d. Find $P(D|C)$.

Exercise:

Problem:

A special deck of cards has ten cards. Four are green, three are blue, and three are red. When a card is picked, its color is recorded. An experiment consists of first picking a card and then tossing a coin.

a. List the sample space.
b. Let *A* be the event that a blue card is picked first, followed by landing a head on the coin toss. Find $P(A)$.
c. Let *B* be the event that a red or green is picked, followed by landing a head on the coin toss. Are the events *A* and *B* mutually exclusive? Explain your answer in one to three complete sentences, including numerical justification.
d. Let *C* be the event that a red or blue is picked, followed by landing a head on the coin toss. Are the events *A* and *C* mutually exclusive? Explain your answer in one to three complete sentences, including numerical justification.

Solution:

Note:
NOTE
The coin toss is independent of the card picked first.

a. {(*G*,*H*) (*G*,*T*) (*B*,*H*) (*B*,*T*) (*R*,*H*) (*R*,*T*)}
b. $P(A) = P(\text{blue})P(\text{head}) = \left(\frac{3}{10}\right)\left(\frac{1}{2}\right) = \frac{3}{20}$
c. Yes, *A* and *B* are mutually exclusive because they cannot happen at the same time; you cannot pick a card that is both blue and also (red or green). $P(A \text{ AND } B) = 0$

d. No, A and C are not mutually exclusive because they can occur at the same time. In fact, C includes all of the outcomes of A; if the card chosen is blue it is also (red or blue). $P(A \text{ AND } C) = P(A) = \frac{3}{20}$, which is not 0.

Bringing It Together

Exercise:

Problem:

A previous year, the weights of the members of the **San Francisco 49ers** and the **Dallas Cowboys** were published in the *San Jose Mercury News*. The factual data are compiled into the following table.

Shirt#	**≤ 210**	**211–250**	**251–290**	**290≤**
1–33	21	5	0	0
34–66	6	18	7	4
66–99	6	12	22	5

For the following, suppose that you randomly select one player from the 49ers or Cowboys.

If having a shirt number from one to 33 and weighing at most 210 pounds were independent events, then what should be true about P(Shirt# 1–33|≤ 210 pounds)?

Exercise:

Problem:

The probability that a male develops some form of cancer in his lifetime is 0.4567. The probability that a male has at least one false positive test result (meaning the test comes back for cancer when the man does not have it) is 0.51. Some of the following questions do not have enough information for you to answer them. Write “not enough information” for those answers. Let C = a man develops cancer in his lifetime and P = man has at least one false positive.

a. $P(C)$ = ______
b. $P(P|C)$ = ______
c. $P(P|C')$ = ______
d. If a test comes up positive, based upon numerical values, can you assume that man has cancer? Justify numerically and explain why or why not.

Solution:

a. $P(C) = 0.4567$
b. not enough information
c. not enough information
d. No, because over half (0.51) of men have at least one false positive text

Exercise:

Problem:

Given events G and H: $P(G) = 0.43$; $P(H) = 0.26$; $P(H \text{ AND } G) =$ 0.14

a. Find $P(H|G)$.
b. Find the probability of the complement of event (H AND G).
c. Find the probability of the complement of event ($H|G$).

Exercise:

Problem:

Given events J and K: $P(J)$ = 0.18; $P(K)$ = 0.37; $P(J \text{ OR } K)$ = 0.45; $P(J|K)$ = 0.27

a. Find $P(J \text{ AND } K)$. (Round to one decimal place.)
b. Find the probability of the complement of event (J AND K).
c. Find the probability of the complement of event (J OR K).

Solution:

a. $P(J|K) = \frac{P(J \text{ AND } K)}{P(K)}$; $0.27 = \frac{P(J \text{ AND } K)}{0.37}$; solve to find $P(J \text{ AND } K)$ = 0.1
b. $P(\text{not}(J \text{ AND } K))$ = 1 - $P(J \text{ AND } K)$ = 1 - 0.1 = 0.9
c. $P(\text{not}(J \text{ OR } K))$ = 1 - $P(J \text{ OR } K)$ = 1 - 0.45 = 0.55

Glossary

Independent Events

The occurrence of one event has no effect on the probability of the occurrence of another event. Events A and B are independent if one of the following is true:

1. $P(A|B) = P(A)$
2. $P(B|A) = P(B)$
3. $P(A \text{ AND } B) = P(A)P(B)$

Mutually Exclusive

Two events are mutually exclusive if the probability that they both happen at the same time is zero. If events A and B are mutually exclusive, then $P(A \text{ AND } B) = 0$.

Dependent Events

If two events are NOT independent, then we say that they are dependent.

Sampling with Replacement

If each member of a population is replaced after it is picked, then that member has the possibility of being chosen more than once.

Sampling without Replacement

When sampling is done without replacement, each member of a population may be chosen only once.

Two Basic Rules of Probability

When calculating probability, there are two rules to consider when determining if two events are independent or dependent and if they are mutually exclusive or not.

The Multiplication Rule

If A and B are two events defined on a **sample space**, then: $P(A \text{ AND } B)$ = $P(B)P(A|B)$.

This rule may also be written as: $P(A|B)$ = $\frac{P(A \text{ AND } B)}{P(B)}$

(The probability of A given B equals the probability of A AND B divided by the probability of B.)

If A and B are **independent**, then $P(A|B)$ = $P(A)$.

Thus, if A and B are **independent**, then

$P(A \text{ AND } B)$ = $P(A|B)P(B)$ becomes $P(A \text{ AND } B)$ = $P(A)P(B)$.

The Addition Rule

If A and B are defined on a sample space, then: $P(A \text{ OR } B)$ = $P(A)$ + $P(B)$ - $P(A \text{ AND } B)$.

If A and B are **mutually exclusive**, then $P(A \text{ AND } B) = 0$.

Thus, if A and B are **mutually exclusive**, then

$P(A \text{ OR } B)$ = $P(A)$ + $P(B)$ - $P(A \text{ AND } B)$ becomes $P(A \text{ OR } B)$ = $P(A)$ + $P(B)$.

Example:

Klaus is trying to choose where to go on vacation. His two choices are: A = New Zealand and B = Alaska

- Klaus can only afford one vacation. The probability that he chooses A is $P(A)$ = 0.6 and the probability that he chooses B is $P(B)$ = 0.35.
- $P(A \text{ AND } B)$ = 0 because Klaus can only afford to take one vacation
- Therefore, the probability that he chooses either New Zealand or Alaska is $P(A \text{ OR } B) = P(A) + P(B)$ = 0.6 + 0.35 = 0.95. Note that the probability that he does not choose to go anywhere on vacation must be 0.05.

Example:
Carlos plays college soccer. He makes a goal 65% of the time he shoots. Carlos is going to attempt two goals in a row in the next game. A = the event Carlos is successful on his first attempt. $P(A)$ = 0.65. B = the event Carlos is successful on his second attempt. $P(B)$ = 0.65. Carlos tends to shoot in streaks. The probability that he makes the second goal **GIVEN** that he made the first goal is 0.90.

Exercise:

Problem: a. What is the probability that he makes both goals?

Solution:

a. The problem is asking you to find $P(A \text{ AND } B) = P(B \text{ AND } A)$. Since $P(B|A)$ = 0.90, $P(B \text{ AND } A) = P(B|A)\ P(A)$ = (0.90)(0.65) = 0.585

The probability that Carlos makes both the first and second goals is 0.585.

Exercise:

Problem:

b. What is the probability that Carlos makes either the first goal or the second goal?

Solution:

b. The problem is asking you to find $P(A \text{ OR } B)$.

$P(A \text{ OR } B) = P(A) + P(B) - P(A \text{ AND } B) = 0.65 + 0.65 - 0.585 = 0.715$

Carlos makes either the first goal or the second goal with probability 0.715.

Exercise:

Problem: c. Are *A* and *B* independent?

Solution:

c. No, they are not, because $P(A \text{ AND } B) = 0.585$.

$P(A)P(B) = (0.65)(0.65) = 0.423$

$0.423 \neq 0.585 = P(A \text{ AND } B)$

So, $P(A \text{ AND } B)$ is **not** equal to $P(A)P(B)$.

Exercise:

Problem: d. Are *A* and *B* mutually exclusive?

Solution:

d. No, they are not because $P(A \text{ AND } B) = 0.585$.

To be mutually exclusive, $P(A \text{ AND } B)$ must equal zero.

Note:
Try It
Exercise:

Problem:

Helen plays basketball. For free throws, she makes the shot 75% of the time. Helen must now attempt two free throws. C = the event that Helen makes the first shot. $P(C)$ = 0.75. D = the event Helen makes the second shot. $P(D)$ = 0.75. The probability that Helen makes the second free throw given that she made the first is 0.85. What is the probability that Helen makes both free throws?

Solution:

$P(D|C) = 0.85$

$P(C \text{ AND } D) = P(D \text{ AND } C)$
$P(D \text{ AND } C) = P(D|C)P(C) = (0.85)(0.75) = 0.6375$
Helen makes the first and second free throws with probability 0.6375.

Example:
A community swim team has **150** members. **Seventy-five** of the members are advanced swimmers. **Forty-seven** of the members are intermediate swimmers. The remainder are novice swimmers. **Forty** of the advanced swimmers practice four times a week. **Thirty** of the intermediate swimmers practice four times a week. **Ten** of the novice swimmers practice four times a week. Suppose one member of the swim team is chosen randomly.

Exercise:

Problem:

a. What is the probability that the member is a novice swimmer?

Solution:

a. $\frac{28}{150}$

Exercise:

Problem:

b. What is the probability that the member practices four times a week?

Solution:

b. $\frac{80}{150}$

Exercise:

Problem:

c. What is the probability that the member is an advanced swimmer and practices four times a week?

Solution:

c. $\frac{40}{150}$

Exercise:

Problem:

d. What is the probability that a member is an advanced swimmer and an intermediate swimmer? Are being an advanced swimmer and an intermediate swimmer mutually exclusive? Why or why not?

Solution:

d. P(advanced AND intermediate) = 0, so these are mutually exclusive events. A swimmer cannot be an advanced swimmer and an intermediate swimmer at the same time.

Exercise:

Problem:

e. Are being a novice swimmer and practicing four times a week independent events? Why or why not?

Solution:

e. No, these are not independent events.

P(novice AND practices four times per week) = 0.0667

P(novice)P(practices four times per week) = 0.0996

$0.0667 \neq 0.0996$

Example:

Felicity attends Modesto JC in Modesto, CA. The probability that Felicity enrolls in a math class is 0.2 and the probability that she enrolls in a speech class is 0.65. The probability that she enrolls in a math class GIVEN that she enrolls in speech class is 0.25.

Let: M = math class, S = speech class, $M|S$ = math given speech

Exercise:

Problem:

a. What is the probability that Felicity enrolls in math and speech? Find $P(M \text{ AND } S) = P(M|S)P(S)$.
b. What is the probability that Felicity enrolls in math or speech classes? Find $P(M \text{ OR } S) = P(M) + P(S) - P(M \text{ AND } S)$.

c. Are M and S independent? Is $P(M|S) = P(M)$?
d. Are M and S mutually exclusive? Is $P(M \text{ AND } S) = 0$?

Solution:

a. 0.1625, b. 0.6875, c. No, d. No

Note:
Try It
Exercise:

Problem:

A student goes to the library. Let events B = the student checks out a book and D = the student check out a DVD. Suppose that $P(B) = 0.40$, $P(D) = 0.30$ and $P(D|B) = 0.5$.

a. Find $P(B \text{ AND } D)$.
b. Find $P(B \text{ OR } D)$.

Solution:

a. $P(B \text{ AND } D) = P(D|B)P(B) = (0.5)(0.4) = 0.20$.
b. $P(B \text{ OR } D) = P(B) + P(D) - P(B \text{ AND } D) = 0.40 + 0.30 - 0.20 = 0.50$

Example:
Studies show that about one woman in seven (approximately 14.3%) who live to be 90 will develop breast cancer. Suppose that of those women who develop breast cancer, a test is negative 2% of the time. Also suppose that in the general population of women, the test for breast cancer is negative about

85% of the time. Let B = woman develops breast cancer and let N = tests negative. Suppose one woman is selected at random.

Exercise:

Problem:

a. What is the probability that the woman develops breast cancer? What is the probability that woman tests negative?

Solution:

a. $P(B) = 0.143$; $P(N) = 0.85$

Exercise:

Problem:

b. Given that the woman has breast cancer, what is the probability that she tests negative?

Solution:

b. $P(N|B) = 0.02$

Exercise:

Problem:

c. What is the probability that the woman has breast cancer AND tests negative?

Solution:

c. $P(B \text{ AND } N) = P(B)P(N|B) = (0.143)(0.02) = 0.0029$

Exercise:

Problem:

d. What is the probability that the woman has breast cancer or tests negative?

Solution:

d. $P(B \text{ OR } N) = P(B) + P(N) - P(B \text{ AND } N) = 0.143 + 0.85 - 0.0029 = 0.9901$

Exercise:

Problem:

e. Are having breast cancer and testing negative independent events?

Solution:

e. No. $P(N) = 0.85$; $P(N|B) = 0.02$. So, $P(N|B)$ does not equal $P(N)$.

Exercise:

Problem:

f. Are having breast cancer and testing negative mutually exclusive?

Solution:

f. No. $P(B \text{ AND } N) = 0.0029$. For B and N to be mutually exclusive, $P(B \text{ AND } N)$ must be zero.

Note:
Try It
Exercise:

Problem:

A school has 200 seniors of whom 140 will be going to college next year. Forty will be going directly to work. The remainder are taking a gap year. Fifty of the seniors going to college play sports. Thirty of the seniors going directly to work play sports. Five of the seniors taking a gap year play sports. What is the probability that a senior is going to college and plays sports?

Solution:

Let A = student is a senior going to college.

Let B = student plays sports.

$P(B) = \frac{140}{200}$

$P(B|A) = \frac{50}{140}$

$P(A \text{ AND } B) = P(B|A)P(A)$

$P(A \text{ AND } B) = \left(\frac{140}{200}\right)\left(\frac{50}{140}\right) = \frac{1}{4}$

Example:
Exercise:

Problem: Refer to the information in [link]. P = tests positive.

a. Given that a woman develops breast cancer, what is the probability that she tests positive. Find $P(P|B) = 1 - P(N|B)$.
b. What is the probability that a woman develops breast cancer and tests positive. Find $P(B \text{ AND } P) = P(P|B)P(B)$.
c. What is the probability that a woman does not develop breast cancer. Find $P(B') = 1 - P(B)$.

d. What is the probability that a woman tests positive for breast cancer. Find $P(P) = 1 - P(N)$.

Solution:

a. 0.98; b. 0.1401; c. 0.857; d. 0.15

Note:
Try It
Exercise:

Problem:

A student goes to the library. Let events B = the student checks out a book and D = the student checks out a DVD. Suppose that $P(B) = 0.40$, $P(D) = 0.30$ and $P(D|B) = 0.5$.

a. Find $P(B')$.
b. Find $P(D \text{ AND } B)$.
c. Find $P(B|D)$.
d. Find $P(D \text{ AND } B')$.
e. Find $P(D|B')$.

Solution:

a. $P(B') = 0.60$
b. $P(D \text{ AND } B) = P(D|B)P(B) = 0.20$
c. $P(B|D) = \frac{P(B \text{ AND } D)}{P(D)} = \frac{(0.20)}{(0.30)} = 0.67$
d. $P(D \text{ AND } B') = P(D) - P(D \text{ AND } B) = 0.30 - 0.20 = 0.10$
e. $P(D|B') = P(D \text{ AND } B')P(B') = (0.10)(0.60) = 0.06$

References

DiCamillo, Mark, Mervin Field. “The File Poll.” Field Research Corporation. Available online at http://www.field.com/fieldpollonline/subscribers/Rls2443.pdf (accessed May 2, 2013).

Rider, David, “Ford support plummeting, poll suggests,” The Star, September 14, 2011. Available online at http://www.thestar.com/news/gta/2011/09/14/ford_support_plummeting_poll_suggests.html (accessed May 2, 2013).

“Mayor’s Approval Down.” News Release by Forum Research Inc. Available online at http://www.forumresearch.com/forms/News Archives/News Releases/74209_TO_Issues_-_Mayoral_Approval_%28Forum_Research%29%2820130320%29.pdf (accessed May 2, 2013).

“Roulette.” Wikipedia. Available online at http://en.wikipedia.org/wiki/Roulette (accessed May 2, 2013).

Shin, Hyon B., Robert A. Kominski. “Language Use in the United States: 2007.” United States Census Bureau. Available online at http://www.census.gov/hhes/socdemo/language/data/acs/ACS-12.pdf (accessed May 2, 2013).

Data from the Baseball-Almanac, 2013. Available online at www.baseball-almanac.com (accessed May 2, 2013).

Data from U.S. Census Bureau.

Data from the Wall Street Journal.

Data from The Roper Center: Public Opinion Archives at the University of Connecticut. Available online at http://www.ropercenter.uconn.edu/ (accessed May 2, 2013).

Data from Field Research Corporation. Available online at www.field.com/fieldpollonline (accessed May 2,2 013).

Section Review

The multiplication rule and the addition rule are used for computing the probability of *A* and *B*, as well as the probability of *A* or *B* for two given events *A*, *B* defined on the sample space. In sampling with replacement each member of a population is replaced after it is picked, so that member has the possibility of being chosen more than once, and the events are considered to be independent. In sampling without replacement, each member of a population may be chosen only once, and the events are considered to be not independent. The events *A* and *B* are mutually exclusive events when they do not have any outcomes in common.

Formula Review

The multiplication rule: $P(A \text{ AND } B)$ = $P(B)P(A|B)$

The addition rule: $P(A \text{ OR } B)$ = $P(A)$ + $P(B)$ - $P(A \text{ AND } B)$

Use the following information to answer the next ten exercises. Forty-eight percent of all Californians registered voters prefer life in prison without parole over the death penalty for a person convicted of first degree murder. Among Latino California registered voters, 55% prefer life in prison without parole over the death penalty for a person convicted of first degree murder. 37.6% of all Californians are Latino.

In this problem, let:

- *C* = Californians (registered voters) preferring life in prison without parole over the death penalty for a person convicted of first degree murder.
- *L* = Latino Californians

Suppose that one Californian is randomly selected.

Exercise:

Problem: Find $P(C)$.

Exercise:

Problem: Find $P(L)$.

Solution:

0.376

Exercise:

Problem: Find $P(C|L)$.

Exercise:

Problem: In words, what is $C|L$?

Solution:

$C|L$ means, given the person chosen is a Latino Californian, the person is a registered voter who prefers life in prison without parole for a person convicted of first degree murder.

Exercise:

Problem: Find $P(L \text{ AND } C)$.

Exercise:

Problem: In words, what is L AND C?

Solution:

L AND C is the event that the person chosen is a Latino California registered voter who prefers life without parole over the death penalty for a person convicted of first degree murder.

Exercise:

Problem: Are L and C independent events? Show why or why not.

Exercise:

Problem: Find $P(L \text{ OR } C)$.

Solution:

0.6492

Exercise:

Problem: In words, what is L OR C?

Exercise:

Problem:

Are L and C mutually exclusive events? Show why or why not.

Solution:

No, because $P(L \text{ AND } C)$ does not equal 0.

Homework

Exercise:

Problem:

On February 28, 2013, a Field Poll Survey reported that 61% of California registered voters approved of allowing two people of the same gender to marry and have regular marriage laws apply to them. Among 18 to 39 year olds (California registered voters), the approval rating was 78%. Six in ten California registered voters said that the upcoming Supreme Court’s ruling about the constitutionality of California’s Proposition 8 was either very or somewhat important to them. Out of those CA registered voters who support same-sex marriage, 75% say the ruling is important to them.

In this problem, let:

- C = California registered voters who support same-sex marriage.
- B = California registered voters who say the Supreme Court's ruling about the constitutionality of California's Proposition 8 is very or somewhat important to them
- A = California registered voters who are 18 to 39 years old.

a. Find $P(C)$.
b. Find $P(B)$.
c. Find $P(C|A)$.
d. Find $P(B|C)$.
e. In words, what is $C|A$?
f. In words, what is $B|C$?
g. Find $P(C \text{ AND } B)$.
h. In words, what is C AND B?
i. Find $P(C \text{ OR } B)$.
j. Are C and B mutually exclusive events? Show why or why not.

Exercise:

Problem:

After Rob Ford, the mayor of Toronto, announced his plans to cut budget costs in late 2011, the Forum Research polled 1,046 people to measure the mayor's popularity. Everyone polled expressed either approval or disapproval. These are the results their poll produced:

- In early 2011, 60 percent of the population approved of Mayor Ford's actions in office.
- In mid-2011, 57 percent of the population approved of his actions.
- In late 2011, the percentage of popular approval was measured at 42 percent.

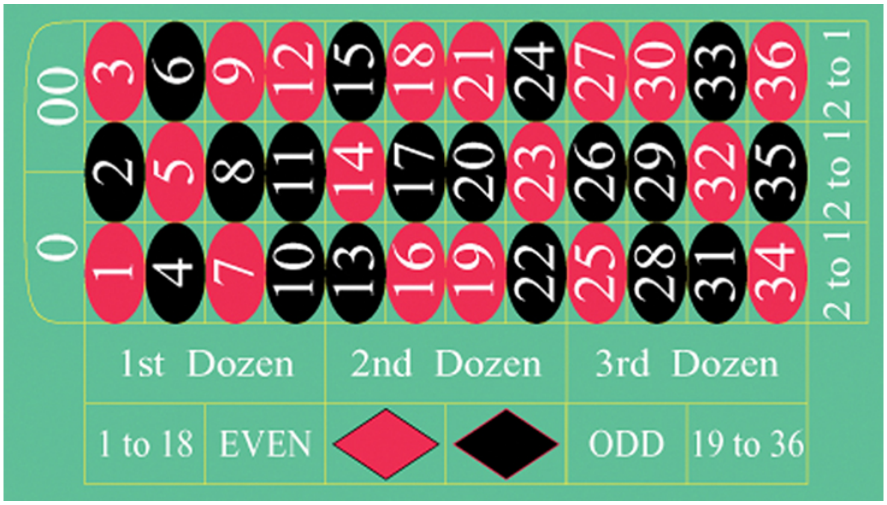
a. What is the sample size for this study?
b. What proportion in the poll disapproved of Mayor Ford, according to the results from late 2011?
c. How many people polled responded that they approved of Mayor Ford in late 2011?

d. What is the probability that a person supported Mayor Ford, based on the data collected in mid-2011?
e. What is the probability that a person supported Mayor Ford, based on the data collected in early 2011?

Solution:

a. The Forum Research surveyed 1,046 Torontonians.
b. 58%
c. 42% of 1,046 = 439 (rounding to the nearest integer)
d. 0.57
e. 0.60.

Use the following information to answer the next three exercises. The casino game, roulette, allows the gambler to bet on the probability of a ball, which spins in the roulette wheel, landing on a particular color, number, or range of numbers. The table used to place bets contains of 38 numbers, and each number is assigned to a color and a range.


00
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
2 to 1
2 to 1
2 to 1
1st Dozen
2nd Dozen
3rd Dozen
1 to 18
EVEN
ODD
19 to 36

(credit: film8ker/wikibooks)

Exercise:

Problem:

a. List the sample space of the 38 possible outcomes in roulette.
b. You bet on red. Find $P(\text{red})$.
c. You bet on -1st 12- (1st Dozen). Find $P(\text{-1st 12-})$.
d. You bet on an even number. Find $P(\text{even number})$.
e. Is getting an odd number the complement of getting an even number? Why?
f. Find two mutually exclusive events.
g. Are the events Even and 1st Dozen independent?

Exercise:

Problem:

Compute the probability of winning the following types of bets:

a. Betting on two lines that touch each other on the table as in 1-2-3-4-5-6
b. Betting on three numbers in a line, as in 1-2-3
c. Betting on one number
d. Betting on four numbers that touch each other to form a square, as in 10-11-13-14
e. Betting on two numbers that touch each other on the table, as in 10-11 or 10-13
f. Betting on 0-00-1-2-3
g. Betting on 0-1-2; or 0-00-2; or 00-2-3

Solution:

a. $P(\text{Betting on two line that touch each other on the table}) = \frac{6}{38}$
b. $P(\text{Betting on three numbers in a line}) = \frac{3}{38}$
c. $P(\text{Betting on one number}) = \frac{1}{38}$
d. $P(\text{Betting on four number that touch each other to form a square}) = \frac{4}{38}$

e. P(Betting on two number that touch each other on the table) = $\frac{2}{38}$
f. P(Betting on 0-00-1-2-3) = $\frac{5}{38}$
g. P(Betting on 0-1-2; or 0-00-2; or 00-2-3) = $\frac{3}{38}$

Exercise:

Problem:

Compute the probability of winning the following types of bets:

a. Betting on a color
b. Betting on one of the dozen groups
c. Betting on the range of numbers from 1 to 18
d. Betting on the range of numbers 19–36
e. Betting on one of the columns
f. Betting on an even or odd number (excluding zero)

Exercise:

Problem:

Suppose that you have eight cards. Five are green and three are yellow. The five green cards are numbered 1, 2, 3, 4, and 5. The three yellow cards are numbered 1, 2, and 3. The cards are well shuffled. You randomly draw one card.

- G = card drawn is green
- E = card drawn is even-numbered

a. List the sample space.
b. $P(G)$ = _____
c. $P(G|E)$ = _____
d. $P(G \text{ AND } E)$ = _____
e. $P(G \text{ OR } E)$ = _____
f. Are G and E mutually exclusive? Justify your answer numerically.

Solution:

a. $\{G1, G2, G3, G4, G5, Y1, Y2, Y3\}$
b. $\frac{5}{8}$
c. $\frac{2}{3}$
d. $\frac{2}{8}$
e. $\frac{6}{8}$
f. No, because $P(G \text{ AND } E)$ does not equal 0.

Exercise:

Problem: Roll two fair dice. Each die has six faces.

a. List the sample space.
b. Let A be the event that either a three or four is rolled first, followed by an even number. Find $P(A)$.
c. Let B be the event that the sum of the two rolls is at most seven. Find $P(B)$.
d. In words, explain what "$P(A|B)$" represents. Find $P(A|B)$.
e. Are A and B mutually exclusive events? Explain your answer in one to three complete sentences, including numerical justification.
f. Are A and B independent events? Explain your answer in one to three complete sentences, including numerical justification.

Exercise:

Problem:

An experiment consists of tossing a nickel, a dime, and a quarter. Of interest is the side the coin lands on.

a. List the sample space.
b. Let A be the event that there are at least two tails. Find $P(A)$.
c. Let B be the event that the first and second tosses land on heads. Are the events A and B mutually exclusive? Explain your answer in one to three complete sentences, including justification.

Solution:

a. S = {(HHH), (HHT), (HTH), (HTT), (THH), (THT), (TTH), (TTT)}
b. $\frac{4}{8}$
c. Yes, because if A has occurred, it is impossible to obtain two tails. In other words, $P(A \text{ AND } B) = 0$.

Exercise:

Problem:

An experiment consists of first rolling a die and then tossing a coin.

a. List the sample space.
b. Let A be the event that either a three or a four is rolled first, followed by landing a head on the coin toss. Find $P(A)$.
c. Let B be the event that a two is rolled and the coin toss lands on heads. Are the events A and B mutually exclusive? Explain your answer in one to three complete sentences, including numerical justification.

Exercise:

Problem: Y and Z are independent events.

a. Rewrite the basic Addition Rule $P(Y \text{ OR } Z) = P(Y) + P(Z) - P(Y \text{ AND } Z)$ using the information that Y and Z are independent events.
b. Use the rewritten rule to find $P(Z)$ if $P(Y \text{ OR } Z) = 0.71$ and $P(Y) = 0.42$.

Solution:

a. If Y and Z are independent, then $P(Y \text{ AND } Z) = P(Y)P(Z)$, so $P(Y \text{ OR } Z) = P(Y) + P(Z) - P(Y)P(Z)$.
b. 0.5

Exercise:

Problem:

G and H are mutually exclusive events. $P(G) = 0.5$, $P(H) = 0.3$

a. Explain why the following statement MUST be false: $P(H|G) = 0.4$.
b. Find $P(H \text{ OR } G)$.
c. Are G and H independent or dependent events? Explain in a complete sentence.

Exercise:

Problem:

Approximately 281,000,000 people over age five live in the United States. Of these people, 55,000,000 speak a language other than English at home. Of those who speak another language at home, 62.3% speak Spanish.

Let: E = speaks English at home; E' = speaks another language at home; S = speaks Spanish;

Finish each probability statement by matching the correct answer.

Probability Statements	**Answers**
a. $P(E') =$	i. 0.8043
b. $P(E) =$	ii. 0.623
c. $P(S \text{ and } E') =$	iii. 0.1957

Probability Statements	Answers
d. $P(S \mid E')$ =	iv. 0.1219

Solution:

iii i iv ii

Exercise:

Problem:

1994, the U.S. government held a lottery to issue 55,000 Green Cards (permits for non-citizens to work legally in the U.S.). Renate Deutsch, from Germany, was one of approximately 6.5 million people who entered this lottery. Let G = won green card.

a. What was Renate's chance of winning a Green Card? Write your answer as a probability statement.
b. In the summer of 1994, Renate received a letter stating she was one of 110,000 finalists chosen. Once the finalists were chosen, assuming that each finalist had an equal chance to win, what was Renate's chance of winning a Green Card? Write your answer as a conditional probability statement. Let F = was a finalist.
c. Are G and F independent or dependent events? Justify your answer numerically and also explain why.
d. Are G and F mutually exclusive events? Justify your answer numerically and explain why.

Exercise:

Problem:

Three professors at George Washington University did an experiment to determine if economists are more selfish than other people. They dropped 64 stamped, addressed envelopes with $10 cash in different classrooms on the George Washington campus. 44% were returned overall. From the economics classes 56% of the envelopes were returned. From the business, psychology, and history classes 31% were returned.

Let: R = money returned; E = economics classes; O = other classes

a. Write a probability statement for the overall percent of money returned.
b. Write a probability statement for the percent of money returned out of the economics classes.
c. Write a probability statement for the percent of money returned out of the other classes.
d. Is money being returned independent of the class? Justify your answer numerically and explain it.
e. Based upon this study, do you think that economists are more selfish than other people? Explain why or why not. Include numbers to justify your answer.

Solution:

a. $P(R) = 0.44$
b. $P(R|E) = 0.56$
c. $P(R|O) = 0.31$
d. No, whether the money is returned is not independent of which class the money was placed in. There are several ways to justify this mathematically, but one is that the money placed in economics classes is not returned at the same overall rate; $P(R|E) \neq P(R)$.
e. No, this study definitely does not support that notion; *in fact*, it suggests the opposite. The money placed in the economics classrooms was returned at a higher rate than the money place in all classes collectively; $P(R|E) > P(R)$.

Exercise:

Problem:

The following table of data obtained from www.baseball-almanac.com shows hit information for four players. Suppose that one hit from the table is randomly selected.

Name	**Single**	**Double**	**Triple**	**Home Run**	**Total Hits**
Babe Ruth	1,517	506	136	714	2,873
Jackie Robinson	1,054	273	54	137	1,518
Ty Cobb	3,603	174	295	114	4,189
Hank Aaron	2,294	624	98	755	3,771
Total	8,471	1,577	583	1,720	12,351

Are "the hit being made by Hank Aaron" and "the hit being a double" independent events?

a. Yes, because $P(\text{hit by Hank Aaron}|\text{hit is a double}) = P(\text{hit by Hank Aaron})$
b. No, because $P(\text{hit by Hank Aaron}|\text{hit is a double}) \neq P(\text{hit is a double})$

c. No, because P(hit is by Hank Aaron|hit is a double) ≠ P(hit by Hank Aaron)
d. Yes, because P(hit is by Hank Aaron|hit is a double) = P(hit is a double)

Exercise:

Problem:

United Blood Services is a blood bank that serves more than 500 hospitals in 18 states. According to their website, a person with type O blood and a negative Rh factor (Rh-) can donate blood to any person with any bloodtype. Their data show that 43% of people have type O blood and 15% of people have Rh- factor; 52% of people have type O or Rh- factor.

a. Find the probability that a person has both type O blood and the Rh- factor.
b. Find the probability that a person does NOT have both type O blood and the Rh- factor.

Solution:

a. P(type O OR Rh-) = P(type O) + P(Rh-) - P(type O AND Rh-)

   0.52 = 0.43 + 0.15 - P(type O AND Rh-); solve to find P(type O AND Rh-) = 0.06

   6% of people have type O, Rh- blood

b. P(NOT(type O AND Rh-)) = 1 - P(type O AND Rh-) = 1 - 0.06 = 0.94

   94% of people do not have type O, Rh- blood

Exercise:

Problem:

At a college, 72% of courses have final exams and 46% of courses require research papers. Suppose that 32% of courses have a research paper and a final exam. Let F be the event that a course has a final exam. Let R be the event that a course requires a research paper.

a. Find the probability that a course has a final exam or a research project.
b. Find the probability that a course has NEITHER of these two requirements.

Exercise:

Problem:

In a box of assorted cookies, 36% contain chocolate and 12% contain nuts. Of those, 8% contain both chocolate and nuts. Sean is allergic to both chocolate and nuts.

a. Find the probability that a cookie contains chocolate or nuts (he can't eat it).
b. Find the probability that a cookie does not contain chocolate or nuts (he can eat it).

Solution:

a. Let C = be the event that the cookie contains chocolate. Let N = the event that the cookie contains nuts.
b. $P(C \text{ OR } N) = P(C) + P(N) - P(C \text{ AND } N) = 0.36 + 0.12 - 0.08 = 0.40$
c. $P(\text{NEITHER chocolate NOR nuts}) = 1 - P(C \text{ OR } N) = 1 - 0.40 = 0.60$

Exercise:

Problem:

A college finds that 10% of students have taken a distance learning class and that 40% of students are part time students. Of the part time students, 20% have taken a distance learning class. Let *D* = event that a student takes a distance learning class and *E* = event that a student is a part time student

a. Find $P(D \text{ AND } E)$.
b. Find $P(E|D)$.
c. Find $P(D \text{ OR } E)$.
d. Using an appropriate test, show whether *D* and *E* are independent.
e. Using an appropriate test, show whether *D* and *E* are mutually exclusive.

Glossary

Independent Events

The occurrence of one event has no effect on the probability of the occurrence of another event. Events *A* and *B* are independent if one of the following is true:

1. $P(A|B) = P(A)$
2. $P(B|A) = P(B)$
3. $P(A \text{ AND } B) = P(A)P(B)$

Mutually Exclusive

Two events are mutually exclusive if the probability that they both happen at the same time is zero. If events *A* and *B* are mutually exclusive, then $P(A \text{ AND } B) = 0$.

Contingency Tables

A **contingency table** provides a way of portraying data that can facilitate calculating probabilities. The table helps in determining conditional probabilities quite easily. The table displays sample values in relation to two different variables that may be dependent or contingent on one another. Later on, we will use contingency tables again, but in another manner.

Example:
Suppose a study of speeding violations and drivers who use cell phones produced the following fictional data:

	Speeding violation in the last year	**No speeding violation in the last year**	**Total**
Cell phone user	25	280	305
Not a cell phone user	45	405	450
Total	70	685	755

The total number of people in the sample is 755. The row totals are 305 and 450. The column totals are 70 and 685. Notice that 305 + 450 = 755 and 70 + 685 = 755.

Calculate the following probabilities using the table.

Exercise:

Problem: a. Find $P(\text{Person is a cell phone user})$.

Solution:

a. $\frac{\text{number of cell phone users}}{\text{total number in study}} = \frac{305}{755}$

Exercise:

Problem: b. Find $P(\text{person had no violation in the last year})$.

Solution:

b. $\frac{\text{number that had no violation}}{\text{total number in study}} = \frac{685}{755}$

Exercise:

Problem: c. Find P(Person had no violation in the last year AND was a cell phone user).

Solution:

c. $\frac{280}{755}$

Exercise:

Problem: d. Find P(Person is a cell phone user OR person had no violation in the last year).

Solution:

d. $\left(\frac{305}{755} + \frac{685}{755}\right) - \frac{280}{755} = \frac{710}{755}$

Exercise:

Problem: e. Find P(Person is a cell phone user GIVEN person had a violation in the last year).

Solution:

e. $\frac{25}{70}$ (The sample space is reduced to the number of persons who had a violation.)

Exercise:

Problem: f. Find P(Person had no violation last year GIVEN person was not a cell phone user)

Solution:

f. $\frac{405}{450}$ (The sample space is reduced to the number of persons who were not cell phone users.)

Note:
Try it
Exercise:

Problem:

The following table shows the number of athletes who stretch before exercising and how many had injuries within the past year.

	Injury in last year	**No injury in last year**	**Total**
Stretches	55	295	350
Does not stretch	231	219	450
Total	286	514	800

a. What is $P(\text{athlete stretches before exercising})$?
b. What is $P(\text{athlete stretches before exercising}|\text{no injury in the last year})$?

Solution:

a. $P(\text{athlete stretches before exercising}) = \frac{350}{800} = 0.4375$
b. $P(\text{athlete stretches before exercising}|\text{no injury in the last year}) = \frac{295}{514} = 0.5739$

Example:
The following table shows a random sample of 100 hikers and the areas of hiking they prefer.

Sex	**The Coastline**	**Near Lakes and Streams**	**On Mountain Peaks**	**Total**
Female	18	16	___	45
Male	___	___	14	55
Total	___	41	___	___

Hiking Area Preference

Exercise:

Problem: a. Complete the table.

Solution:

a.

Sex	**The Coastline**	**Near Lakes and Streams**	**On Mountain Peaks**	**Total**
Female	18	16	**11**	45
Male	**16**	**25**	14	55
Total	**34**	41	**25**	**100**

Hiking Area Preference

Exercise:

Problem: b. Are the events "being female" and "preferring the coastline" independent events?

Let F = being female and let C = preferring the coastline.

1. Find $P(F \text{ AND } C)$.
2. Find $P(F)P(C)$

Are these two numbers the same? If they are, then F and C are independent. If they are not, then F and C are not independent.

Solution:

b.

1. $P(F \text{ AND } C) = \frac{18}{100} = 0.18$
2. $P(F)P(C) = \left(\frac{45}{100}\right)\left(\frac{34}{100}\right) = (0.45)(0.34) = 0.153$

$P(F \text{ AND } C) \neq P(F)P(C)$, so the events F and C are not independent.

Exercise:

Problem:

c. Find the probability that a person is male given that the person prefers hiking near lakes and streams. Let M = being male, and let L = prefers hiking near lakes and streams.

1. What word tells you this is a conditional?
2. Fill in the blanks and calculate the probability: P (___|___) = ___.
3. Is the sample space for this problem all 100 hikers? If not, what is it?

Solution:

c.

1. The word 'given' tells you that this is a conditional.
2. $P(M|L) = \frac{25}{41}$
3. No, the sample space for this problem is the 41 hikers who prefer lakes and streams.

Exercise:

Problem:

d. Find the probability that a person is female or prefers hiking on mountain peaks. Let *F* = being female, and let *P* = prefers mountain peaks.

1. Find $P(F)$.
2. Find $P(P)$.
3. Find $P(F \text{ AND } P)$.
4. Find $P(F \text{ OR } P)$.

Solution:

d.

1. $P(F) = \frac{45}{100}$
2. $P(P) = \frac{25}{100}$
3. $P(F \text{ AND } P) = \frac{11}{100}$
4. $P(F \text{ OR } P) = \frac{45}{100} + \frac{25}{100} - \frac{11}{100} = \frac{59}{100}$

Note:
Try It
Exercise:

Problem:

The following table shows a random sample of 200 cyclists and the routes they prefer. Let *M* = males and *H* = hilly path.

Gender	**Lake Path**	**Hilly Path**	**Wooded Path**	**Total**
Female	45	38	27	110
Male	26	52	12	90
Total	71	90	39	200

a. Out of the males, what is the probability that the cyclist prefers a hilly path?
b. Are the events “being male” and “preferring the hilly path” independent events?

Solution:

a. $P(H|M) = \frac{52}{90} = 0.5778$
b. For M and H to be independent, show $P(H|M) = P(H)$

$P(H|M) = 0.5778$, $P(H) = \frac{90}{200} = 0.45$

$P(H|M)$ does not equal $P(H)$ so M and H are NOT independent.

Example:
Muddy Mouse lives in a cage with three doors. If Muddy goes out the first door, the probability that he gets caught by Alissa the cat is $\frac{1}{5}$ and the probability he is not caught is $\frac{4}{5}$. If he goes out the second door, the probability he gets caught by Alissa is $\frac{1}{4}$ and the probability he is not caught is $\frac{3}{4}$. The probability that Alissa catches Muddy coming out of the third door is $\frac{1}{2}$ and the probability she does not catch Muddy is $\frac{1}{2}$. It is equally likely that Muddy will choose any of the three doors so the probability of choosing each door is $\frac{1}{3}$.

Caught or Not	Door One	Door Two	Door Three	Total
Caught	$\frac{1}{15}$	$\frac{1}{12}$	$\frac{1}{6}$	____
Not Caught	$\frac{4}{15}$	$\frac{3}{12}$	$\frac{1}{6}$	____
Total	____	____	____	1

Door Choice

- The first entry $\frac{1}{15} = \left(\frac{1}{5}\right)\left(\frac{1}{3}\right)$ is $P(\text{Door One AND Caught})$
- The entry $\frac{4}{15} = \left(\frac{4}{5}\right)\left(\frac{1}{3}\right)$ is $P(\text{Door One AND Not Caught})$

Verify the remaining entries.

Exercise:

Problem:

a. Complete the probability contingency table. Calculate the entries for the totals. Verify that the lower-right corner entry is 1.

Solution:

a.

Caught or Not	Door One	Door Two	Door Three	Total
Caught	$\frac{1}{15}$	$\frac{1}{12}$	$\frac{1}{6}$	$\frac{19}{60}$
Not Caught	$\frac{4}{15}$	$\frac{3}{12}$	$\frac{1}{6}$	$\frac{41}{60}$
Total	$\frac{5}{15}$	$\frac{4}{12}$	$\frac{2}{6}$	1

Door Choice

Exercise:

Problem: b. What is the probability that Alissa does not catch Muddy?

Solution:

b. $\frac{41}{60}$

Exercise:

Problem:

c. What is the probability that Muddy chooses Door One OR Door Two given that Muddy is caught by Alissa?

Solution:

c. $\frac{9}{19}$

Example:
The following table contains the number of crimes per 100,000 inhabitants from 2008 to 2011 in the U.S.

Year	Robbery	Burglary	Rape	Vehicle	Total
2008	145.7	732.1	29.7	314.7	

Year	**Robbery**	**Burglary**	**Rape**	**Vehicle**	**Total**
2009	133.1	717.7	29.1	259.2	
2010	119.3	701	27.7	239.1	
2011	113.7	702.2	26.8	229.6	
Total					

United States Crime Index Rates Per 100,000 Inhabitants 2008–2011

Exercise:

Problem: TOTAL each column and each row. Total data = 4,520.7

a. Find $P(\text{2009 AND Robbery})$.
b. Find $P(\text{2010 AND Burglary})$.
c. Find $P(\text{2010 OR Burglary})$.
d. Find $P(\text{2011}|\text{Rape})$.
e. Find $P(\text{Vehicle}|\text{2008})$.

Solution:

a. 0.0294, b. 0.1551, c. 0.7165, d. 0.2365, e. 0.2575

Note:
Try It
Exercise:

Problem:

The following table relates the weights and heights of a group of individuals participating in an observational study.

Weight/Height	**Tall**	**Medium**	**Short**	**Totals**
Obese	18	28	14	
Normal	20	51	28	
Underweight	12	25	9	

Weight/Height	Tall	Medium	Short	Totals
Totals				

a. Find the total for each row and column
b. Find the probability that a randomly chosen individual from this group is Tall.
c. Find the probability that a randomly chosen individual from this group is Obese and Tall.
d. Find the probability that a randomly chosen individual from this group is Tall given that the idividual is Obese.
e. Find the probability that a randomly chosen individual from this group is Obese given that the individual is Tall.
f. Find the probability a randomly chosen individual from this group is Tall and Underweight.
g. Are the events Obese and Tall independent?

Solution:

Weight/Height	Tall	Medium	Short	Totals
Obese	18	28	14	60
Normal	20	51	28	99
Underweight	12	25	9	46
Totals	50	104	51	205

a. Row Totals: 60, 99, 46. Column totals: 50, 104, 51.
b. $P(\text{Tall}) = \frac{50}{205} = 0.244$
c. $P(\text{Obese AND Tall}) = \frac{18}{205} = 0.088$
d. $P(\text{Tall}|\text{Obese}) = \frac{18}{60} = 0.3$
e. $P(\text{Obese}|\text{Tall}) = \frac{18}{50} = 0.36$
f. $P(\text{Tall AND Underweight}) = \frac{12}{205} = 0.0585$
g. No. $P(\text{Tall})$ does not equal $P(\text{Tall}|\text{Obese})$.

References

“Blood Types.” American Red Cross, 2013. Available online at http://www.redcrossblood.org/learn-about-blood/blood-types (accessed May 3, 2013).

Data from the National Center for Health Statistics, part of the United States Department of Health and Human Services.

Data from United States Senate. Available online at www.senate.gov (accessed May 2, 2013).

Haiman, Christopher A., Daniel O. Stram, Lynn R. Wilkens, Malcom C. Pike, Laurence N. Kolonel, Brien E. Henderson, and Loīc Le Marchand. “Ethnic and Racial Differences in the Smoking-Related Risk of Lung Cancer.” The New England Journal of Medicine, 2013. Available online at http://www.nejm.org/doi/full/10.1056/NEJMoa033250 (accessed May 2, 2013).

“Human Blood Types.” Unite Blood Services, 2011. Available online at http://www.unitedbloodservices.org/learnMore.aspx (accessed May 2, 2013).

Samuel, T. M. “Strange Facts about RH Negative Blood.” eHow Health, 2013. Available online at http://www.ehow.com/facts_5552003_strange-rh-negative-blood.html (accessed May 2, 2013).

“United States: Uniform Crime Report – State Statistics from 1960–2011.” The Disaster Center. Available online at http://www.disastercenter.com/crime/ (accessed May 2, 2013).

Section Review

There are several tools you can use to help organize and sort data when calculating probabilities. Contingency tables help display data and are particularly useful when calculating probabilites that have multiple dependent variables.

Use the following information to answer the next four exercises. The table below shows a random sample of musicians and how they learned to play their instruments.

Gender	Self-taught	Studied in School	Private Instruction	Total
Female	12	38	22	72
Male	19	24	15	58
Total	31	62	37	130

Exercise:

Problem: Find $P(\text{musician is a female})$.

Exercise:

Problem: Find $P(\text{musician is a male AND had private instruction})$.

Solution:

$P(\text{musician is a male AND had private instruction})$ = $\frac{15}{130}$ = $\frac{3}{26}$ = 0.12

Exercise:

Problem: Find $P(\text{musician is a female OR is self taught})$.

Exercise:

Problem:

Are the events “being a female musician” and “learning music in school” mutually exclusive events?

Solution:

$P(\text{being a female musician AND learning music in school})$ = $\frac{38}{130}$ = $\frac{19}{65}$ = 0.29

$P(\text{being a female musician})P(\text{learning music in school})$ = $\left(\frac{72}{130}\right)\left(\frac{62}{130}\right)$ = $\frac{4{,}464}{16{,}900} = \frac{1{,}116}{4{,}225}$ = 0.26

No, they are not independent because $P(\text{being a female musician AND learning music in school})$ is not equal to $P(\text{being a female musician})P(\text{learning music in school})$.

Bringing It Together

Use the following information to answer the next seven exercises. An article in the *New England Journal of Medicine*, reported about a study of smokers in California and Hawaii. In one part of the report, the self-reported ethnicity and smoking levels per day were given. Of the people smoking at most ten cigarettes per day, there were 9,886 African Americans, 2,745 Native Hawaiians, 12,831 Latinos, 8,378 Japanese Americans, and 7,650 Whites. Of the people smoking 11 to 20 cigarettes per day, there were 6,514 African Americans, 3,062 Native Hawaiians, 4,932 Latinos, 10,680 Japanese Americans, and 9,877 Whites. Of the people smoking 21 to 30 cigarettes per day, there were 1,671 African Americans, 1,419 Native Hawaiians, 1,406 Latinos, 4,715 Japanese Americans, and 6,062 Whites. Of the people smoking at least 31 cigarettes per day, there were 759 African Americans, 788 Native Hawaiians, 800 Latinos, 2,305 Japanese Americans, and 3,970 Whites.

Exercise:

Problem:

Complete the table using the data provided. Suppose that one person from the study is randomly selected. Find the probability that person smoked 11 to 20 cigarettes per day.

Smoking Level	**African American**	**Native Hawaiian**	**Latino**	**Japanese Americans**	**White**	**TOTALS**
1–10						
11–20						

Smoking Level	**African American**	**Native Hawaiian**	**Latino**	**Japanese Americans**	**White**	**TOTALS**
21–30						
31+						
TOTALS						

Smoking Levels by Ethnicity

Exercise:

Problem:

Suppose that one person from the study is randomly selected. Find the probability that person smoked 11 to 20 cigarettes per day.

Solution:

$\frac{35{,}065}{100{,}450}$

Exercise:

Problem: Find the probability that the person was Latino.

Exercise:

Problem:

In words, explain what it means to pick one person from the study who is “Japanese American **AND** smokes 21 to 30 cigarettes per day.” Also, find the probability.

Solution:

To pick one person from the study who is Japanese American AND smokes 21 to 30 cigarettes per day means that the person has to meet both criteria: both Japanese American and smokes 21 to 30 cigarettes. The sample space should include everyone in the study. The probability is $\frac{4{,}715}{100{,}450}$.

Exercise:

Problem:

In words, explain what it means to pick one person from the study who is “Japanese American **OR** smokes 21 to 30 cigarettes per day.” Also, find the probability.

Exercise:

Problem:

In words, explain what it means to pick one person from the study who is “Japanese American **GIVEN** that person smokes 21 to 30 cigarettes per day.” Also, find the probability.

Solution:

To pick one person from the study who is Japanese American given that person smokes 21-30 cigarettes per day, means that the person must fulfill both criteria and the sample space is reduced to those who smoke 21-30 cigarettes per day. The probability is $\frac{4715}{15,273}$.

Exercise:

Problem: Prove that smoking level/day and ethnicity are dependent events.

Homework

Use the information in the the following table to answer the next eight exercises. The table shows the political party affiliation of each of 67 members of the US Senate in June 2012, and when they are up for reelection.

Up for reelection:	**Democratic Party**	**Republican Party**	**Other**	**Total**
November 2014	20	13	0	
November 2016	10	24	0	
Total				

Exercise:

Problem: What is the probability that a randomly selected senator has an “Other” affiliation?

Solution:

0

Exercise:

Problem:

What is the probability that a randomly selected senator is up for reelection in November 2016?

Exercise:

Problem:

What is the probability that a randomly selected senator is a Democrat and up for reelection in November 2016?

Solution:

$\frac{10}{67}$

Exercise:

Problem:

What is the probability that a randomly selected senator is a Republican or is up for reelection in November 2014?

Exercise:

Problem:

Suppose that a member of the US Senate is randomly selected. Given that the randomly selected senator is up for reelection in November 2016, what is the probability that this senator is a Democrat?

Solution:

$\frac{10}{34}$

Exercise:

Problem:

Suppose that a member of the US Senate is randomly selected. What is the probability that the senator is up for reelection in November 2014, knowing that this senator is a Republican?

Exercise:

Problem: The events “Republican” and “Up for reelection in 2016” are ________

a. mutually exclusive.
b. independent.
c. both mutually exclusive and independent.
d. neither mutually exclusive nor independent.

Solution:

d

Exercise:

Problem: The events “Other” and “Up for reelection in November 2016” are ________

a. mutually exclusive.
b. independent.
c. both mutually exclusive and independent.
d. neither mutually exclusive nor independent.

Exercise:

Problem:

The following table gives the number of suicides estimated in the U.S. for a recent year by age, race (black or white), and sex. We are interested in possible relationships between age, race, and sex. We will let suicide victims be our population.

Race and Sex	**1–14**	**15–24**	**25–64**	**over 64**	**TOTALS**
white, male	210	3,360	13,610		22,050
white, female	80	580	3,380		4,930
black, male	10	460	1,060		1,670
black, female	0	40	270		330
all others					
TOTALS	310	4,650	18,780		29,760

a. Fill in the column for the suicides for individuals over age 64.
b. Fill in the row for all other races.
c. Find the probability that a randomly selected individual was a white male.
d. Find the probability that a randomly selected individual was a black female.
e. Find the probability that a randomly selected individual was black
f. Find the probability that a randomly selected individual was a black or white male.
g. Out of the individuals over age 64, find the probability that a randomly selected individual was a black or white male.

Solution:

a.

Race and Sex	**1–14**	**15–24**	**25–64**	**over 64**	**TOTALS**
white, male	210	3,360	13,610	4,870	22,050
white, female	80	580	3,380	890	4,930
black, male	10	460	1,060	140	1,670
black, female	0	40	270	20	330
all others				100	
TOTALS	310	4,650	18,780	6,020	29,760

b.

Race and Sex	**1–14**	**15–24**	**25–64**	**over 64**	**TOTALS**
white, male	210	3,360	13,610	4,870	22,050
white, female	80	580	3,380	890	4,930
black, male	10	460	1,060	140	1,670
black, female	0	40	270	20	330
all others	10	210	460	100	780
TOTALS	310	4,650	18,780	6,020	29,760

c. $\frac{22{,}050}{29{,}760}$
d. $\frac{330}{29{,}760}$
e. $\frac{2{,}000}{29{,}760}$
f. $\frac{23{,}720}{29{,}760}$
g. $\frac{5{,}010}{6{,}020}$

Use the following information to answer the next two exercises. The table of data obtained from *www.baseball-almanac.com* shows hit information for four well known baseball players. Suppose that one hit from the table is randomly selected.

NAME	**Single**	**Double**	**Triple**	**Home Run**	**TOTAL HITS**
Babe Ruth	1,517	506	136	714	2,873
Jackie Robinson	1,054	273	54	137	1,518
Ty Cobb	3,603	174	295	114	4,189
Hank Aaron	2,294	624	98	755	3,771
TOTAL	8,471	1,577	583	1,720	12,351

Exercise:

Problem: Find $P(\text{hit was made by Babe Ruth})$.

a. $\frac{1518}{2873}$

b. $\frac{2873}{12351}$
c. $\frac{583}{12351}$
d. $\frac{4189}{12351}$

Exercise:

Problem: Find $P(\text{hit was made by Ty Cobb}|\text{The hit was a Home Run})$.

a. $\frac{4189}{12351}$
b. $\frac{114}{1720}$
c. $\frac{1720}{4189}$
d. $\frac{114}{12351}$

Solution:

b

Exercise:

Problem:

The following table identifies a group of children by one of four hair colors, and by type of hair.

Hair Type	**Brown**	**Blond**	**Black**	**Red**	**Totals**
Wavy	20		15	3	43
Straight	80	15		12	
Totals		20			215

a. Complete the table.
b. What is the probability that a randomly selected child will have wavy hair?
c. What is the probability that a randomly selected child will have either brown or blond hair?
d. What is the probability that a randomly selected child will have wavy brown hair?
e. What is the probability that a randomly selected child will have red hair, given that he or she has straight hair?
f. If *B* is the event of a child having brown hair, find the probability of the complement of *B*.
g. In words, what does the complement of *B* represent?

Exercise:

Problem:

In a previous year, the weights of the members of the **San Francisco 49ers** and the **Dallas Cowboys** were published in the *San Jose Mercury News*. The factual data were compiled into the following table.

Shirt#	**≤ 210**	**211–250**	**251–290**	**> 290**
1–33	21	5	0	0
34–66	6	18	7	4
66–99	6	12	22	5

For the following, suppose that you randomly select one player from the 49ers or Cowboys.

a. Find the probability that his shirt number is from 1 to 33.
b. Find the probability that he weighs at most 210 pounds.
c. Find the probability that his shirt number is from 1 to 33 AND he weighs at most 210 pounds.
d. Find the probability that his shirt number is from 1 to 33 OR he weighs at most 210 pounds.
e. Find the probability that his shirt number is from 1 to 33 GIVEN that he weighs at most 210 pounds.

Solution:

a. $\frac{26}{106}$
b. $\frac{33}{106}$
c. $\frac{21}{106}$
d. $\left(\frac{26}{106}\right) + \left(\frac{33}{106}\right) - \left(\frac{21}{106}\right) = \left(\frac{38}{106}\right)$
e. $\frac{21}{33}$

Glossary

contingency table
: the method of displaying a frequency distribution as a table with rows and columns to show how two variables may be dependent (contingent) upon each other; the table provides an easy way to calculate conditional probabilities.

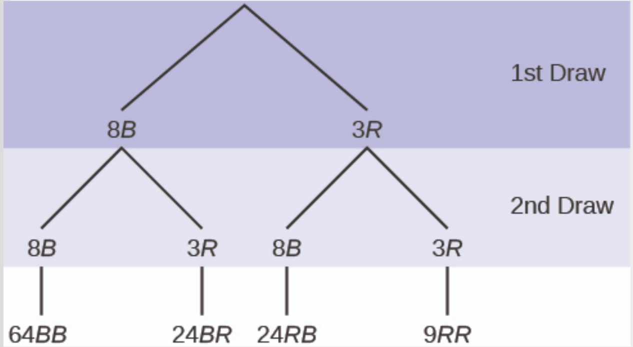
Tree and Venn Diagrams (Optional)

Sometimes, when the probability problems are complex, it can be helpful to graph the situation. Tree diagrams and Venn diagrams are two tools that can be used to visualize and solve conditional probabilities.

Tree Diagrams

A **tree diagram** is a special type of graph used to determine the outcomes of an experiment. It consists of "branches" that are labeled with either frequencies or probabilities. Tree diagrams can make some probability problems easier to visualize and solve. The following example illustrates how to use a tree diagram.

Example:
In an urn, there are 11 balls. Three balls are red (*R*) and eight balls are blue (*B*). Draw two balls, one at a time, **with replacement**. "With replacement" means that you put the first ball back in the urn before you select the second ball. The tree diagram using frequencies that show all the possible outcomes follows.



Total = 64 + 24 + 24 + 9 = 121

The first set of branches represents the first draw. The second set of branches represents the second draw. Each of the outcomes is distinct. In fact, we can list each red ball as *R*1, *R*2, and *R*3 and each blue ball as *B*1, *B*2, *B*3, *B*4, *B*5, *B*6, *B*7, and *B*8. Then the nine *RR* outcomes can be written as:

*R*1*R*1 *R*1*R*2 *R*1*R*3 *R*2*R*1 *R*2*R*2 *R*2*R*3 *R*3*R*1 *R*3*R*2 *R*3*R*3
The other outcomes are similar.
There are a total of 11 balls in the urn. Draw two balls, one at a time, with replacement. There are 11(11) = 121 outcomes, the size of the **sample space**.

Exercise:

Problem: a. List the 24 *BR* outcomes: *B*1*R*1, *B*1*R*2, *B*1*R*3, ...

Solution:

a. $B1R1$ $B1R2$ $B1R3$ $B2R1$ $B2R2$ $B2R3$ $B3R1$ $B3R2$ $B3R3$ $B4R1$ $B4R2$ $B4R3$ $B5R1$ $B5R2$ $B5R3$ $B6R1$ $B6R2$ $B6R3$ $B7R1$ $B7R2$ $B7R3$ $B8R1$ $B8R2$ $B8R3$

Exercise:

Problem: b. Using the tree diagram, calculate $P(RR)$.

Solution:

b. $P(RR) = \left(\frac{3}{11}\right)\left(\frac{3}{11}\right) = \frac{9}{121}$

Exercise:

Problem: c. Using the tree diagram, calculate $P(RB \text{ OR } BR)$.

Solution:

c. $P(RB \text{ OR } BR) = \left(\frac{3}{11}\right)\left(\frac{8}{11}\right) + \left(\frac{8}{11}\right)\left(\frac{3}{11}\right) = \frac{48}{121}$

Exercise:

Problem: d. Using the tree diagram, calculate $P(R \text{ on 1st draw AND } B \text{ on 2nd draw})$.

Solution:

d. $P(R \text{ on 1st draw AND } B \text{ on 2nd draw}) = P(RB) = \left(\frac{3}{11}\right)\left(\frac{8}{11}\right) = \frac{24}{121}$

Exercise:

Problem: e. Using the tree diagram, calculate $P(R \text{ on 2nd draw GIVEN } B \text{ on 1st draw})$.

Solution:

e. $P(R \text{ on 2nd draw GIVEN } B \text{ on 1st draw}) = P(R \text{ on 2nd} | B \text{ on 1st}) = \frac{24}{88} = \frac{3}{11}$

This problem is a conditional one. The sample space has been reduced to those outcomes that already have a blue on the first draw. There are 24 + 64 = 88 possible outcomes (24 BR and 64 BB). Twenty-four of the 88 possible outcomes are BR. $\frac{24}{88} = \frac{3}{11}$.

Exercise:

Problem: f. Using the tree diagram, calculate $P(BB)$.

Solution:

f. $P(BB) = \frac{64}{121}$

Exercise:

Problem:

g. Using the tree diagram, calculate $P(B \text{ on the 2nd draw given } R \text{ on the first draw})$.

Solution:

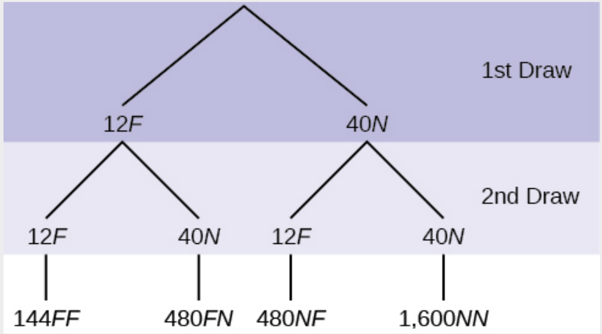
g. $P(B \text{ on 2nd draw}|R \text{ on 1st draw}) = \frac{8}{11}$

There are 9 + 24 outcomes that have *R* on the first draw (9 *RR* and 24 *RB*). The sample space is then 9 + 24 = 33. 24 of the 33 outcomes have *B* on the second draw. The probability is then $\frac{24}{33}$.

Note:
Try It
Exercise:

Problem:

In a standard deck, there are 52 cards. 12 cards are face cards (event *F*) and 40 cards are not face cards (event *N*). Draw two cards, one at a time, with replacement. All possible outcomes are shown in the tree diagram as frequencies. Using the tree diagram, calculate $P(FF)$.



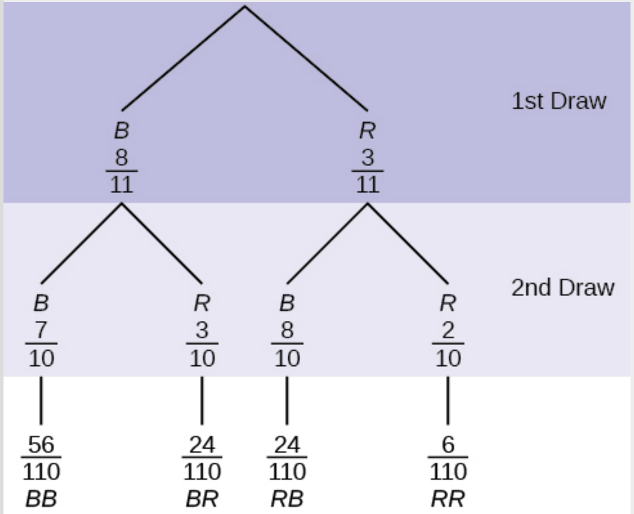
Solution:

Total number of outcomes is 144 + 480 + 480 + 1600 = 2,704.

$$P(FF) = \frac{144}{144 + 480 + 480 + 1{,}600} = \frac{144}{2{,}704} = \frac{9}{169}$$

Example:

An urn has three red marbles and eight blue marbles in it. Draw two marbles, one at a time, this time without replacement, from the urn. **"Without replacement"** means that you do not put the first ball back before you select the second marble. Following is a tree diagram for this situation. The branches are labeled with probabilities instead of frequencies. The numbers at the ends of the branches are calculated by multiplying the numbers on the two corresponding branches, for example, $\left(\frac{3}{11}\right)\left(\frac{2}{10}\right) = \frac{6}{110}$.



$$\text{Total} = \frac{56+24+24+6}{110} = \frac{110}{110} = 1$$

Note:

NOTE

If you draw a red on the first draw from the three red possibilities, there are two red marbles left to draw on the second draw. You do not put back or replace the first marble after you have drawn it. You draw **without replacement**, so that on the second draw there are ten marbles left in the urn.

Calculate the following probabilities using the tree diagram.

Exercise:

Problem: a. $P(RR)$ = ________

Solution:

a. $P(RR) = \left(\frac{3}{11}\right)\left(\frac{2}{10}\right) = \frac{6}{110}$

Exercise:

Problem: b. Fill in the blanks:

$P(RB \text{ OR } BR) = \left(\frac{3}{11}\right)\left(\frac{8}{10}\right) + (\_\_\_)(\_\_\_) = \frac{48}{110}$

Solution:

b. $P(RB \text{ OR } BR) = \left(\frac{3}{11}\right)\left(\frac{8}{10}\right) + \left(\frac{8}{11}\right)\left(\frac{3}{10}\right) = \frac{48}{110}$

Exercise:

Problem: c. $P(R \text{ on 2nd} | B \text{ on 1st})$ =

Solution:

c. $P(R \text{ on 2nd} | B \text{ on 1st}) = \frac{3}{10}$

Exercise:

Problem: d. Fill in the blanks.

$P(R \text{ on 1st AND } B \text{ on 2nd}) = P(RB) = (\_\_\_)(\_\_\_) = \frac{24}{100}$

Solution:

d. $P(R \text{ on 1st AND } B \text{ on 2nd}) = P(RB) = \left(\frac{3}{11}\right)\left(\frac{8}{10}\right) = \frac{24}{100}$

Exercise:

Problem: e. Find $P(BB)$.

Solution:

e. $P(BB) = \left(\frac{8}{11}\right)\left(\frac{7}{10}\right)$

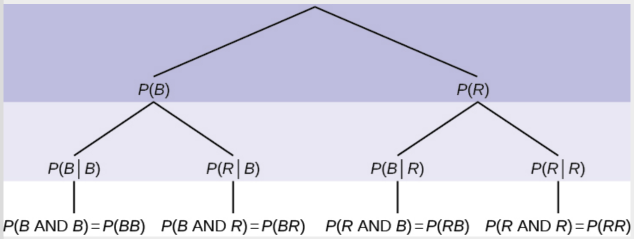
Exercise:

Problem: f. Find $P(B \text{ on 2nd}|R \text{ on 1st})$.

Solution:

f. Using the tree diagram, $P(B \text{ on 2nd}|R \text{ on 1st}) = P(R\,|B) = \frac{8}{10}$.

If we are using probabilities, we can label the tree in the following general way.

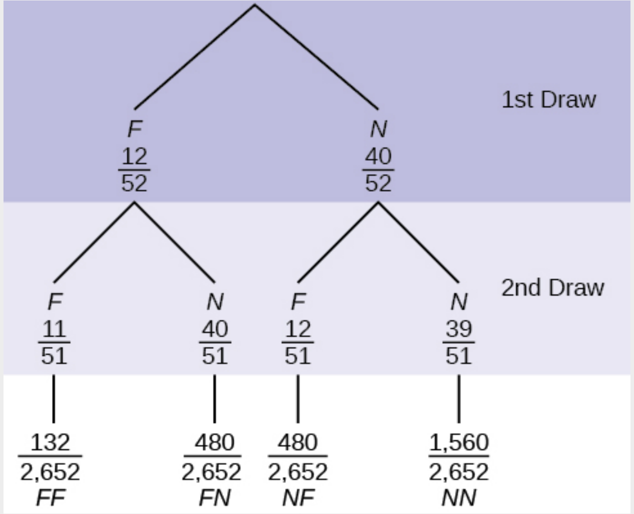

P(B)
P(R)
P(B|B)
P(R|B)
P(B|R)
P(R|R)
P(B AND B) = P(BB)
P(B AND R) = P(BR)
P(R AND B) = P(RB)
P(R AND R) = P(RR)

- $P(R\,|R)$ here means $P(R \text{ on 2nd}|R \text{ on 1st})$
- $P(B\,|R)$ here means $P(B \text{ on 2nd}|R \text{ on 1st})$
- $P(R\,|B)$ here means $P(R \text{ on 2nd}|B \text{ on 1st})$
- $P(B\,|B)$ here means $P(B \text{ on 2nd}|B \text{ on 1st})$

Note:
Try It
Exercise:

Problem:

In a standard deck, there are 52 cards. Twelve cards are face cards (*F*) and 40 cards are not face cards (*N*). Draw two cards, one at a time, without replacement. The tree diagram is labeled with all possible probabilities.

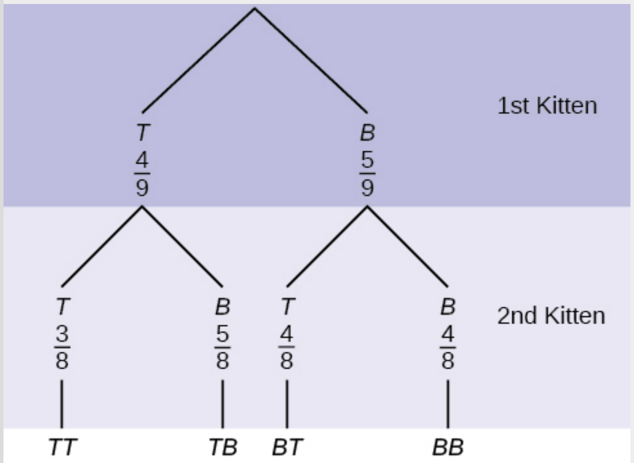

1st Draw
F
12/52
N
40/52
2nd Draw
F
11/51
N
40/51
F
12/51
N
39/51
132/2,652
FF
480/2,652
FN
480/2,652
NF
1,560/2,652
NN

a. Find $P(FN \text{ OR } NF)$.
b. Find $P(N|F)$.
c. Find $P(\text{at most one face card})$.
Hint: "At most one face card" means zero or one face card.
d. Find $P(\text{at least on face card})$.
Hint: "At least one face card" means one or two face cards.

Solution:

a. $P(FN \text{ OR } NF) = \frac{480}{2,652} + \frac{480}{2,652} = \frac{960}{2,652} = \frac{80}{221}$
b. $P(N|F) = \frac{40}{51}$
c. $P(\text{at most one face card}) = \frac{(480 + 480 + 1,560)}{2,652} = \frac{2,520}{2,652}$
d. $P(\text{at least one face card}) = \frac{(132 + 480 + 480)}{2,652} = \frac{1,092}{2,652}$

Example:
A litter of kittens available for adoption at the Humane Society has four tabby kittens and five black kittens. A family comes in and randomly selects two kittens (without replacement) for adoption.



Exercise:

Problem:

a. What is the probability that both kittens are tabby?

a.$\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)$ b.$\left(\frac{4}{9}\right)\left(\frac{4}{9}\right)$ c.$\left(\frac{4}{9}\right)\left(\frac{3}{8}\right)$ d.$\left(\frac{4}{9}\right)\left(\frac{5}{9}\right)$

b. What is the probability that one kitten of each coloring is selected?

a.$\left(\frac{4}{9}\right)\left(\frac{5}{9}\right)$ b.$\left(\frac{4}{9}\right)\left(\frac{5}{8}\right)$ c.$\left(\frac{4}{9}\right)\left(\frac{5}{9}\right)+\left(\frac{5}{9}\right)\left(\frac{4}{9}\right)$ d.$\left(\frac{4}{9}\right)\left(\frac{5}{8}\right)+\left(\frac{5}{9}\right)\left(\frac{4}{8}\right)$

c. What is the probability that a tabby is chosen as the second kitten when a black kitten was chosen as the first?
d. What is the probability of choosing two kittens of the same color?

Solution:

a. c, b. d, c. $\frac{4}{8}$, d. $\frac{32}{72}$

Note:
Try It
Exercise:

Problem:

Suppose there are four red balls and three yellow balls in a box. Two balls are drawn from the box without replacement. What is the probability that one ball of each coloring is selected?

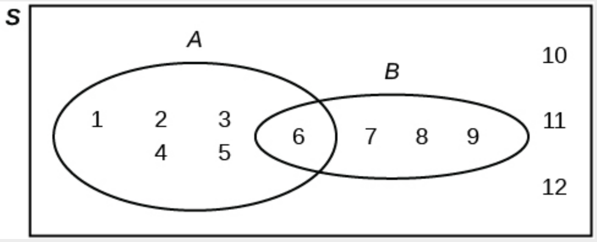
Solution:

$\left(\frac{4}{7}\right)\left(\frac{3}{6}\right) + \left(\frac{3}{7}\right)\left(\frac{4}{6}\right) = \frac{4}{7}$

Venn Diagram

A **Venn diagram** is a picture that represents the outcomes of an experiment. It generally consists of a box that represents the sample space S together with circles or ovals. The circles or ovals represent events.

Example:
Suppose an experiment has the outcomes 1, 2, 3, ... , 12 where each outcome has an equal chance of occurring. Let event *A* = {1, 2, 3, 4, 5, 6} and event *B* = {6, 7, 8, 9}. Then *A* AND *B* = {6} and *A* OR *B* = {1, 2, 3, 4, 5, 6, 7, 8, 9}. The Venn diagram is as follows:

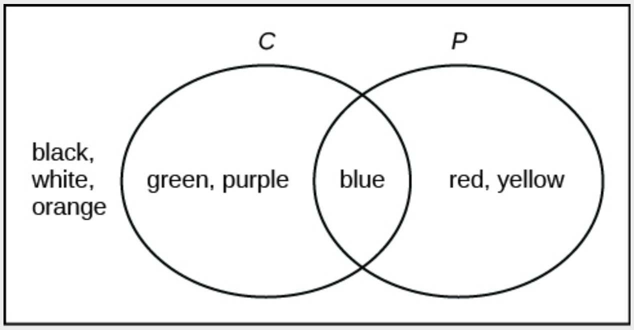


Note:
Try It
Exercise:

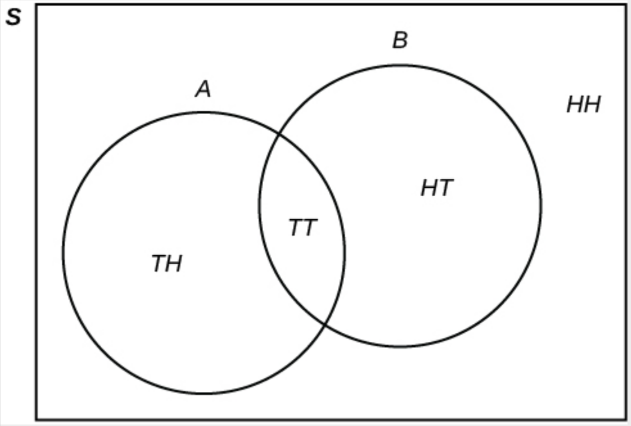
Problem:

Suppose an experiment has outcomes black, white, red, orange, yellow, green, blue, and purple, where each outcome has an equal chance of occurring. Let event C = {green, blue, purple} and event P = {red, yellow, blue}. Then C AND P = {blue} and C OR P = {green, blue, purple, red, yellow}. Draw a Venn diagram representing this situation.

Solution:



Example:
Flip two fair coins. Let A = tails on the first coin. Let B = tails on the second coin. Then A = {TT, TH} and B = {TT, HT}. Therefore, A AND B = {TT}. A OR B = {TH, TT, HT}.
The sample space when you flip two fair coins is X = {HH, HT, TH, TT}. The outcome HH is in NEITHER A NOR B. The Venn diagram is as follows:

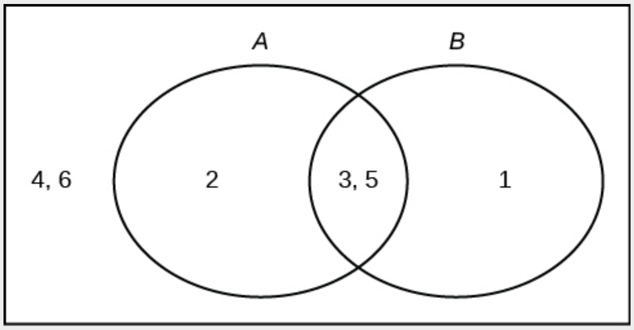


Note:
Try It
Exercise:

Problem:

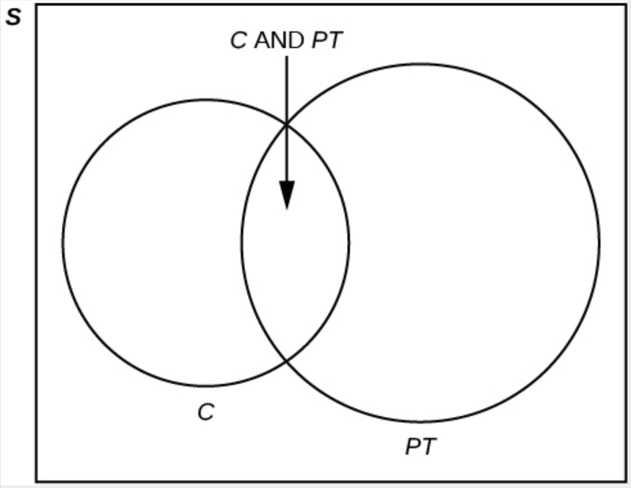
Roll a fair, six-sided die. Let A = a prime number of dots is rolled. Let B = an odd number of dots is rolled. Then A = {2, 3, 5} and B = {1, 3, 5}. Therefore, A AND B = {3, 5}. A OR B = {1, 2, 3, 5}. The sample space for rolling a fair die is S = {1, 2, 3, 4, 5, 6}. Draw a Venn diagram representing this situation.

Solution:



Example:

Forty percent of the students at a local college belong to a club and **50%** work part time. **Five percent** of the students work part time and belong to a club. Draw a Venn diagram showing the relationships. Let C = student belongs to a club and PT = student works part time.



If a student is selected at random, find

- the probability that the student belongs to a club. $P(C)$ = 0.40
- the probability that the student works part time. $P(PT)$ = 0.50
- the probability that the student belongs to a club AND works part time. $P(C \text{ AND } PT)$ = 0.05
- the probability that the student belongs to a club **given** that the student works part time. $P(C|PT) = \frac{P(C \text{ AND } PT)}{P(PT)} = \frac{0.05}{0.50} = 0.1$

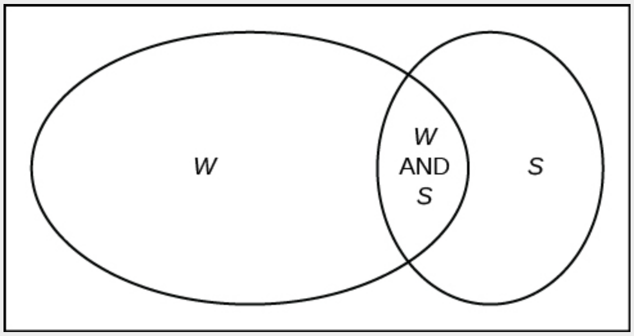
- the probability that the student belongs to a club **OR** works part time. $P(C \text{ OR } PT) = P(C) + P(PT) - P(C \text{ AND } PT) = 0.40 + 0.50 - 0.05 = 0.85$

Note:
Try It
Exercise:

Problem:

Fifty percent of the workers at a factory work a second job, 25% have a spouse who also works, 5% work a second job and have a spouse who also works. Draw a Venn diagram showing the relationships. Let *W* = works a second job and *S* = spouse also works.

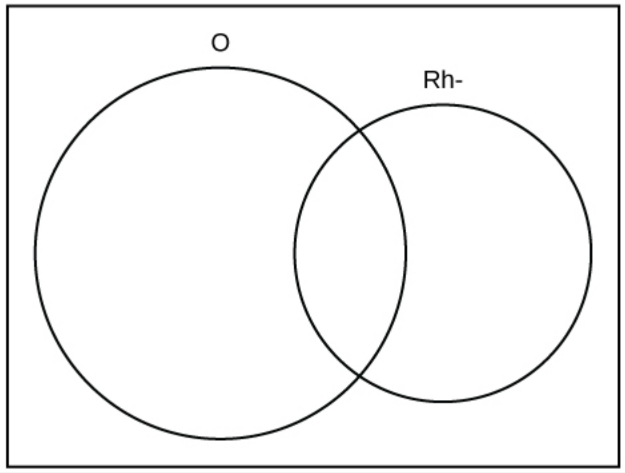
Solution:



Example:
Exercise:

Problem:

A person with type O blood and a negative Rh factor (Rh-) can donate blood to any person with any blood type. Four percent of African Americans have type O blood and a negative RH factor, 5−10% of African Americans have the Rh- factor, and 51% have type O blood.



The “O” circle represents the African Americans with type O blood. The “Rh-“ oval represents the African Americans with the Rh- factor.

We will take the average of 5% and 10% and use 7.5% as the percent of African Americans who have the Rh- factor. Let *O* = African American with Type O blood and *R* = African American with Rh- factor.

a. $P(O)$ = ___________
b. $P(R)$ = ___________
c. $P(O\text{ AND }R)$ = ___________
d. $P(O\text{ OR }R)$ = ____________
e. In the Venn Diagram, describe the overlapping area using a complete sentence.
f. In the Venn Diagram, describe the area in the rectangle but outside both the circle and the oval using a complete sentence.

Solution:

a. 0.51; b. 0.075; c. 0.04; d. 0.545; e. The area represents the African Americans that have type O blood and the Rh- factor. f. The area represents the African Americans that have neither type O blood nor the Rh- factor.

Note:
Try It
Exercise:

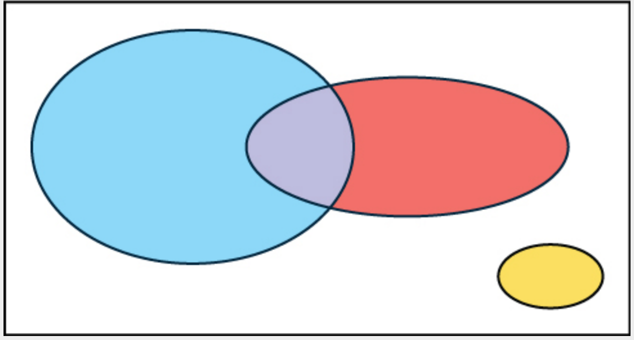
Problem:

In a bookstore, the probability that the customer buys a novel is 0.6, and the probability that the customer buys a non-fiction book is 0.4. Suppose that the probability that the customer buys both is 0.2.

a. Draw a Venn diagram representing the situation.
b. Find the probability that the customer buys either a novel or anon-fiction book.
c. In the Venn diagram, describe the overlapping area using a complete sentence.
d. Suppose that some customers buy only compact disks. Draw an oval in your Venn diagram representing this event.

Solution:

a. and d. In the following Venn diagram below, the blue oval represent customers buying a novel, the red oval represents customer buying non-fiction, and the yellow oval customer who buy compact disks.



b. $P(\text{novel or non-fiction}) = P(\text{Blue OR Red}) = P(\text{Blue}) + P(\text{Red}) - P(\text{Blue AND Red}) = 0.6 + 0.4 - 0.2 = 0.8$.
c. The overlapping area of the blue oval and red oval represents the customers buying both a novel and a nonfiction book.

References

Data from Clara County Public H.D.

Data from the American Cancer Society.

Data from The Data and Story Library, 1996. Available online at http://lib.stat.cmu.edu/DASL/ (accessed May 2, 2013).

Data from the Federal Highway Administration, part of the United States Department of Transportation.

Data from the United States Census Bureau, part of the United States Department of Commerce.

Data from USA Today.

“Environment.” The World Bank, 2013. Available online at http://data.worldbank.org/topic/environment (accessed May 2, 2013).

“Search for Datasets.” Roper Center: Public Opinion Archives, University of Connecticut., 2013. Available online at http://www.ropercenter.uconn.edu/data_access/data/search_for_datasets.html (accessed May 2, 2013).

Section Review

A tree diagram use branches to show the different outcomes of experiments and makes complex probability questions easy to visualize.

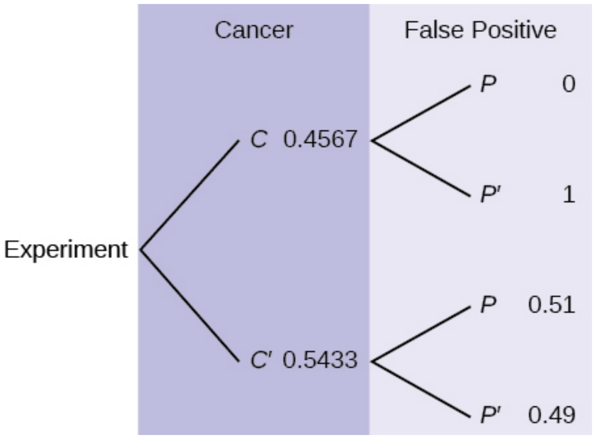
A Venn diagram is a picture that represents the outcomes of an experiment. It generally consists of a box that represents the sample space *S* together with circles or ovals. The circles or ovals represent events. A Venn diagram is especially helpful for visualizing the OR event, the AND event, and the complement of an event and for understanding conditional probabilities.

Exercise:

Problem:

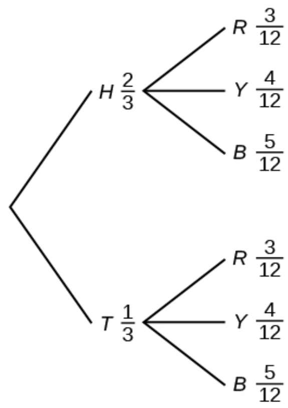
The probability that a man develops some form of cancer in his lifetime is 0.4567. The probability that a man has at least one false positive test result (meaning the test comes back for cancer when the man does not have it) is 0.51. Let: *C* = a man develops cancer in his lifetime; *P* = man has at least one false positive. Construct a tree diagram of the situation.

Solution:



Homework

Use the following information to answer the next two exercises. This tree diagram shows the tossing of an unfair coin followed by drawing one bead from a cup containing three red (*R*), four yellow (*Y*) and five blue (*B*) beads. For the coin, $P(H) = \frac{2}{3}$ and $P(T) = \frac{1}{3}$ where *H* is heads and *T* is tails.


R $\frac{3}{12}$
H $\frac{2}{3}$
Y $\frac{4}{12}$
B $\frac{5}{12}$
R $\frac{3}{12}$
T $\frac{1}{3}$
Y $\frac{4}{12}$
B $\frac{5}{12}$

Exercise:

Problem: Find $P(\text{tossing a Head on the coin AND a Red bead})$

a. $\frac{2}{3}$
b. $\frac{5}{15}$
c. $\frac{6}{36}$
d. $\frac{5}{36}$

Exercise:

Problem: Find $P(\text{Blue bead})$.

a. $\frac{15}{36}$
b. $\frac{10}{36}$
c. $\frac{10}{12}$
d. $\frac{6}{36}$

Solution:

a

Exercise:

Problem:

A box of cookies contains three chocolate and seven butter cookies. Miguel randomly selects a cookie and eats it. Then he randomly selects another cookie and eats it. (How many cookies did he take?)

a. Draw the tree that represents the possibilities for the cookie selections. Write the probabilities along each branch of the tree.
b. Are the probabilities for the flavor of the SECOND cookie that Miguel selects independent of his first selection? Explain.

c. For each complete path through the tree, write the event it represents and find the probabilities.
d. Let S be the event that both cookies selected were the same flavor. Find $P(S)$.
e. Let T be the event that the cookies selected were different flavors. Find $P(T)$ by two different methods: by using the complement rule and by using the branches of the tree. Your answers should be the same with both methods.
f. Let U be the event that the second cookie selected is a butter cookie. Find $P(U)$.

Bringing It Together

Use the following information to answer the next two exercises. Suppose that you have eight cards. Five are green and three are yellow. The cards are well shuffled.

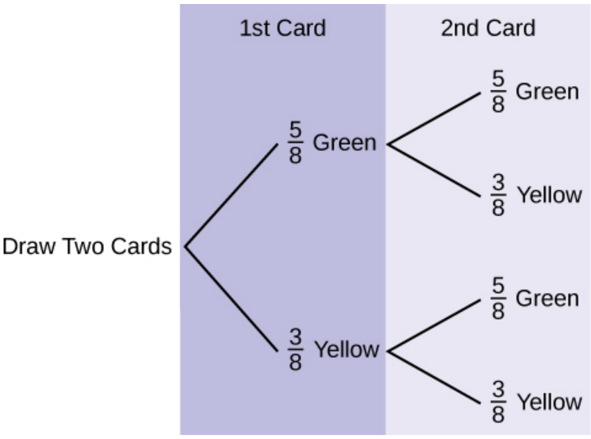
Exercise:

Problem: Suppose that you randomly draw two cards, one at a time, **with replacement**.
Let G_1 = first card is green
Let G_2 = second card is green

a. Draw a tree diagram of the situation.
b. Find $P(G_1 \text{ AND } G_2)$.
c. Find $P(\text{at least one green})$.
d. Find $P(G_2|G_1)$.
e. Are G_2 and G_1 independent events? Explain why or why not.

Solution:

a.



b. $P(GG) = \left(\frac{5}{8}\right)\left(\frac{5}{8}\right) = \frac{25}{64}$
c. $P(\text{at least one green}) = P(GG) + P(GY) + P(YG) = \frac{25}{64} + \frac{15}{64} + \frac{15}{64} = \frac{55}{64}$
d. $P(G\,|G) = \frac{5}{8}$
e. Yes, they are independent because the first card is placed back in the bag before the second card is drawn; the composition of cards in the bag remains the same from draw one to draw two.

Exercise:

Problem: Suppose that you randomly draw two cards, one at a time, **without replacement**.
G_1 = first card is green
G_2 = second card is green

a. Draw a tree diagram of the situation.
b. Find $P(G_1 \text{ AND } G_2)$.
c. Find $P(\text{at least one green})$.
d. Find $P(G_2 | G_1)$.
e. Are G_2 and G_1 independent events? Explain why or why not.

Use the following information to answer the next two exercises. The percent of licensed U.S. drivers (from a recent year) that are female is 48.60. Of the females, 5.03% are age 19 and under; 81.36% are age 20–64; 13.61% are age 65 or over. Of the licensed U.S. male drivers, 5.04% are age 19 and under; 81.43% are age 20–64; 13.53% are age 65 or over.

Exercise:

Problem: Complete the following.

a. Construct a table or a tree diagram of the situation.
b. Find $P($driver is female).
c. Find $P($driver is age 65 or over|driver is female).
d. Find $P($driver is age 65 or over AND female).
e. In words, explain the difference between the probabilities in part c and part d.
f. Find $P($driver is age 65 or over).
g. Are being age 65 or over and being female mutually exclusive events? How do you know?

Solution:

a.

	<20	**20–64**	**>64**	**Totals**
Female	0.0244	0.3954	0.0661	0.486
Male	0.0259	0.4186	0.0695	0.514
Totals	0.0503	0.8140	0.1356	1

b. $P(F)$ = 0.486
c. $P(>64|F)$ = 0.1361
d. $P(>64 \text{ and } F) = P(F)\ P(>64|F)$ = (0.486)(0.1361) = 0.0661
e. $P(>64|F)$ is the percentage of female drivers who are 65 or older and $P(>64 \text{ and } F)$ is the percentage of drivers who are female and 65 or older.
f. $P(>64) = P(>64 \text{ and } F) + P(>64 \text{ and } M)$ = 0.1356
g. No, being female and 65 or older are not mutually exclusive because they can occur at the same time $P(>64 \text{ and } F)$ = 0.0661.

Exercise:

Problem: Suppose that 10,000 U.S. licensed drivers are randomly selected.

a. How many would you expect to be male?
b. Using the table or tree diagram, construct a contingency table of gender versus age group.
c. Using the contingency table, find the probability that out of the age 20–64 group, a randomly selected driver is female.

Exercise:

Problem:

Approximately 86.5% of Americans commute to work by car, truck, or van. Out of that group, 84.6% drive alone and 15.4% drive in a carpool. Approximately 3.9% walk to work and approximately 5.3% take public transportation.

a. Construct a table or a tree diagram of the situation. Include a branch for all other modes of transportation to work.
b. Assuming that the walkers walk alone, what percent of all commuters travel alone to work?
c. Suppose that 1,000 workers are randomly selected. How many would you expect to travel alone to work?
d. Suppose that 1,000 workers are randomly selected. How many would you expect to drive in a carpool?

Solution:

a.

	Car, Truck or Van	**Walk**	**Public Transportation**	**Other**	**Totals**

	Car, Truck or Van	**Walk**	**Public Transportation**	**Other**	**Totals**
Alone	0.7318				
Not Alone	0.1332				
Totals	0.8650	0.0390	0.0530	0.0430	1

b. If we assume that all walkers are alone and that none from the other two groups travel alone (which is a big assumption) we have: $P(\text{Alone})$ = 0.7318 + 0.0390 = 0.7708.
c. Make the same assumptions as in (b) we have: (0.7708)(1,000) = 771
d. (0.1332)(1,000) = 133

Exercise:

Problem:

When the Euro coin was introduced in 2002, two math professors had their statistics students test whether the Belgian one Euro coin was a fair coin. They spun the coin rather than tossing it and found that out of 250 spins, 140 showed a head (event H) while 110 showed a tail (event T). On that basis, they claimed that it is not a fair coin.

a. Based on the given data, find $P(H)$ and $P(T)$.
b. Use a tree to find the probabilities of each possible outcome for the experiment of tossing the coin twice.
c. Use the tree to find the probability of obtaining exactly one head in two tosses of the coin.
d. Use the tree to find the probability of obtaining at least one head.

Exercise:

Problem:

Use the following information to answer the next two exercises. The following are real data from Santa Clara County, CA. As of a certain time, there had been a total of 3,059 documented cases of AIDS in the county. They were grouped into the following categories:

	Homosexual/Bisexual	**IV Drug User***	**Heterosexual Contact**	**Other**	**Totals**
Female	0	70	136	49	____
Male	2,146	463	60	135	____
Totals	____	____	____	____	____

* includes homosexual/bisexual IV drug users

Suppose a person with AIDS in Santa Clara County is randomly selected.

a. Find P(Person is female).
b. Find P(Person has a risk factor heterosexual contact).
c. Find P(Person is female OR has a risk factor of IV drug user).
d. Find P(Person is female AND has a risk factor of homosexual/bisexual).
e. Find P(Person is male AND has a risk factor of IV drug user).
f. Find P(Person is female GIVEN person got the disease from heterosexual contact).
g. Construct a Venn diagram. Make one group females and the other group heterosexual contact.

Solution:

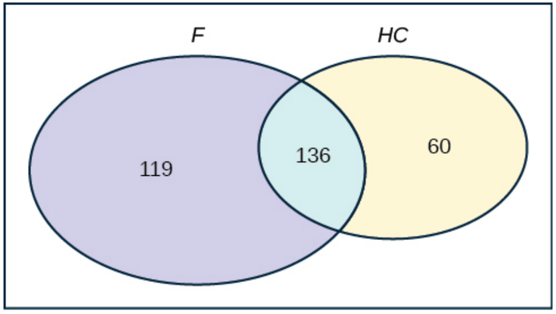
The completed contingency table is as follows:

	Homosexual/Bisexual	**IV Drug User***	**Heterosexual Contact**	**Other**	**Totals**
Female	0	70	136	49	**255**
Male	2,146	463	60	135	**2,804**
Totals	**2,146**	**533**	**196**	**184**	**3,059**

* includes homosexual/bisexual IV drug users

a. $\frac{255}{3059}$
b. $\frac{196}{3059}$

c. $\frac{718}{3059}$
d. 0
e. $\frac{463}{3059}$
f. $\frac{136}{196}$
g.



Exercise:

Problem:

Answer these questions using probability rules. Do NOT use the contingency table. Three thousand fifty-nine cases of AIDS had been reported in Santa Clara County, CA, through a certain date. Those cases will be our population. Of those cases, 6.4% obtained the disease through heterosexual contact and 7.4% are female. Out of the females with the disease, 53.3% got the disease from heterosexual contact.

a. Find $P(\text{Person is female})$.
b. Find $P(\text{Person obtained the disease through heterosexual contact})$.
c. Find $P(\text{Person is female GIVEN person got the disease from heterosexual contact})$
d. Construct a Venn diagram representing this situation. Make one group females and the other group heterosexual contact. Fill in all values as probabilities.

Glossary

Tree Diagram	the useful visual representation of a sample space and events in the form of a “tree” with branches marked by possible outcomes together with associated probabilities (frequencies, relative frequencies)
Venn Diagram	the visual representation of a sample space and events in the form of circles or ovals showing their intersections

Discrete Random Variables: Introduction

class="introduction"

Removed objective regarding geometric and hypergeometric distributions

You can use probability and discrete random variables to calculate the likelihood of lightning striking the ground five times during a half-hour thunderstorm. (Credit: Leszek Leszczynski)

Note:
Chapter Objectives
By the end of this chapter, the student should be able to:

- Recognize and understand discrete probability distribution functions, in general.
- Calculate and interpret expected values.
- Recognize the binomial probability distribution and apply it appropriately.
- Recognize the Poisson probability distribution and apply it appropriately.
- Classify discrete word problems by their distributions.

A student takes a ten-question, true-false quiz. Because the student had such a busy schedule, he or she could not study and guesses randomly at each answer. What is the probability of the student passing the test with at least a 70%?

Small companies might be interested in the number of long-distance phone calls their employees make during the peak time of the day. Suppose the average is 20 calls. What is the probability that the employees make more than 20 long-distance phone calls during the peak time?

These two examples illustrate two different types of probability problems involving discrete random variables. Recall that discrete data are data that you can count. A **random variable** describes the outcomes of a statistical experiment in words. The values of a random variable can vary with each repetition of an experiment.

Random Variable Notation

Upper case letters such as X or Y denote a random variable. Lower case letters like x or y denote the value of a random variable.
If X is a random variable, then X is written in words, and x is given as a number.

For example, let X = the number of heads you get when you toss three fair coins. The sample space for the toss of three fair coins is *TTT*; *THH*; *HTH*; *HHT*; *HTT*; *THT*; *TTH*; *HHH*. Then, x = 0, 1, 2, 3. X is in words and x is a number. Notice that for this example, the x values are countable outcomes. Because you can count the possible values that X can take on and the outcomes are random (the x values 0, 1, 2, 3), X is a discrete random variable.

Note:
Collaborative Exercise
Toss a coin ten times and record the number of heads. After all members of the class have completed the experiment (tossed a coin ten times and counted the number of heads), fill in [link]. Let X = the number of heads in ten tosses of the coin.

x	Frequency of x	Relative Frequency of x

a. Which value(s) of x occurred most frequently?
b. If you tossed the coin 1,000 times, what values could x take on? Which value(s) of x do you think would occur most frequently?
c. What does the relative frequency column sum to?

Glossary

Random Variable (RV)

a characteristic of interest in a population being studied; common notation for variables are upper case Latin letters X, Y, Z,...; common notation for a specific value from the domain (set of all possible values of a variable) are lower case Latin letters x,y, and z. For example, if X is the number of children in a family, then x represents a specific integer 0, 1, 2, 3,.... Variables in statistics differ from variables in intermediate algebra in the two following ways.

- The domain of the random variable (RV) is not necessarily a numerical set; the domain may be expressed in words; for example, if X = hair color then the domain is {black, blond, gray, green, orange}.
- We can tell what specific value x the random variable X takes only after performing the experiment.

Probability Distribution Function (PDF) for a Discrete Random Variable

A discrete **probability distribution function** has two characteristics:

1. Each probability is between zero and one, inclusive.
2. The sum of the probabilities is one.

Example:
A child psychologist is interested in the number of times a newborn baby's crying wakes its mother after midnight. For a random sample of 50 mothers, the following information was obtained. Let X = the number of times per week a newborn baby's crying wakes its mother after midnight. For this example, x = 0, 1, 2, 3, 4, 5.

$P(x)$ = probability that X takes on a value x.

x	***P*(*x*)**
0	$P(x = 0) = \frac{2}{50}$
1	$P(x = 1) = \frac{11}{50}$
2	$P(x = 2) = \frac{23}{50}$
3	$P(x = 3) = \frac{9}{50}$
4	$P(x = 4) = \frac{4}{50}$
5	$P(x = 5) = \frac{1}{50}$

X takes on the values 0, 1, 2, 3, 4, 5. This is a discrete PDF because:

a. Each $P(x)$ is between zero and one, inclusive.
b. The sum of the probabilities is one, that is,

Equation:

$$\frac{2}{50} + \frac{11}{50} + \frac{23}{50} + \frac{9}{50} + \frac{4}{50} + \frac{1}{50} = 1$$

Note:
Try It
Exercise:

Problem:

A hospital researcher is interested in the number of times the average post-op patient will ring the nurse during a 12-hour shift. For a random sample of 50 patients, the following information was obtained. Let X = the number of times a patient rings the nurse during a 12-hour shift. For this exercise, x = 0, 1, 2, 3, 4, 5.

$P(x)$ = the probability that X takes on value x. Why is this a discrete probability distribution function (two reasons)?

x	P**(*x*)**
0	$P(x = 0) = \frac{4}{50}$
1	$P(x = 1) = \frac{8}{50}$

x	$P(x)$
2	$P(x=2)=\frac{16}{50}$
3	$P(x=3)=\frac{14}{50}$
4	$P(x=4)=\frac{6}{50}$
5	$P(x=5)=\frac{2}{50}$

Solution:

Each $P(x)$ is between 0 and 1, inclusive, and the sum of the probabilities is 1, that is:

$$\frac{4}{50}+\frac{8}{50}+\frac{16}{50}+\frac{14}{50}+\frac{6}{50}+\frac{2}{50}=1$$

Example:

Suppose Nancy has classes **three days** a week. She attends classes three days a week **80%** of the time, **two days 15%** of the time, **one day 4%** of the time, and **no days 1%** of the time. Suppose one week is randomly selected.

Exercise:

Problem:

a. Let X = the number of days Nancy ______________________.

Solution:

a. Let X = the number of days Nancy attends class per week.

Exercise:

Problem: b. X takes on what values?

Solution:

b. 0, 1, 2, and 3

Exercise:

Problem:

c. Suppose one week is randomly chosen. Construct a probability distribution table (called a PDF table) like the one in [link]. The table should have two columns labeled x and $P(x)$. What does the $P(x)$ column sum to?

Solution:

c.

x	P**(*x*)**
0	0.01
1	0.04
2	0.15
3	0.80

Note:
Try It
Exercise:

Problem:

Jeremiah has basketball practice two days a week. Ninety percent of the time, he attends both practices. Eight percent of the time, he attends one practice. Two percent of the time, he does not attend either practice. What is X and what values does it take on?

Solution:

X is the number of days Jeremiah attends basketball practice per week. X takes on the values 0, 1, and 2.

Section Review

The characteristics of a probability distribution function (PDF) for a discrete random variable are as follows:

1. Each probability is between zero and one, inclusive (*inclusive* means to include zero and one).
2. The sum of the probabilities is one.

Use the following information to answer the next five exercises: A company wants to evaluate its attrition rate, in other words, how long new hires stay with the company. Over the years, they have established the following probability distribution.

Let X = the number of years a new hire will stay with the company.

Let $P(x)$ = the probability that a new hire will stay with the company x years.

Exercise:

Problem: Complete the following table using the data provided.

x	$P(x)$
0	0.12
1	0.18
2	0.30
3	0.15
4	
5	0.10
6	0.05

Solution:

x	$P(x)$

x	P(*x*)
0	0.12
1	0.18
2	0.30
3	0.15
4	0.10
5	0.10
6	0.05

Exercise:

Problem: $P(x = 4)$ = _______

Exercise:

Problem: $P(x \geq 5)$ = _______

Solution:

0.10 + 0.05 = 0.15

Exercise:

Problem:

On average, how long would you expect a new hire to stay with the company?

Exercise:

Problem: What does the column “$P(x)$” sum to?

Solution:

1

Use the following information to answer the next six exercises: A baker is deciding how many batches of muffins to make to sell in his bakery. He wants to make enough to sell every one and no fewer. Through observation, the baker has established a probability distribution.

x	$P(x)$
1	0.15
2	0.35
3	0.40
4	0.10

Exercise:

Problem: Define the random variable X.

Exercise:

Problem:

What is the probability the baker will sell more than one batch?
$P(x > 1)$ = _______

Solution:

0.35 + 0.40 + 0.10 = 0.85

Exercise:

Problem:

What is the probability the baker will sell exactly one batch?
$P(x = 1)$ = _______

Exercise:

Problem: On average, how many batches should the baker make?

Solution:

1(0.15) + 2(0.35) + 3(0.40) + 4(0.10) = 0.15 + 0.70 + 1.20 + 0.40 = 2.45

Use the following information to answer the next four exercises: Ellen has music practice three days a week. She practices for all of the three days 85% of the time, two days 8% of the time, one day 4% of the time, and no days 3% of the time. One week is selected at random.

Exercise:

Problem: Define the random variable X.

Exercise:

Problem: Construct a probability distribution table for the data.

Solution:

x	P(***x***)
0	0.03
1	0.04
2	0.08
3	0.85

Exercise:

Problem:

We know that for a probability distribution function to be discrete, it must have two characteristics. One is that the sum of the probabilities is one. What is the other characteristic?

Use the following information to answer the next five exercises: Javier volunteers in community events each month. He does not do more than five events in a month. He attends exactly five events 35% of the time, four events 25% of the time, three events 20% of the time, two events 10% of the time, one event 5% of the time, and no events 5% of the time.

Exercise:

Problem: Define the random variable X.

Solution:

Let X = the number of events Javier volunteers for each month.

Exercise:

Problem: What values does X take on?

Exercise:

Problem: Construct a PDF table.

Solution:

x	P**(*x*)**
0	0.05
1	0.05
2	0.10
3	0.20
4	0.25
5	0.35

Exercise:

Problem:

Find the probability that Javier volunteers for less than three events each month. $P(x < 3)$ = _______

Exercise:

Problem:

Find the probability that Javier volunteers for at least one event each month. $P(x > 0)$ = _______

Solution:

1 – 0.05 = 0.95

Homework

Exercise:

Problem:

Suppose that the PDF for the number of years it takes to earn a Bachelor of Science (B.S.) degree is given in the following table.

x	***P*(*x*)**
3	0.05
4	0.40
5	0.30
6	0.15
7	0.10

a. In words, define the random variable X.

b. What does it mean that the values zero, one, and two are not included for x in the PDF?

Glossary

Probability Distribution Function (PDF)

a mathematical description of a discrete random variable (*RV*), given either in the form of an equation (formula) or in the form of a table listing all the possible outcomes of an experiment and the probability associated with each outcome.

Mean or Expected Value and Standard Deviation

The **expected value** is often referred to as the **"long-term" average or mean**. This means that over the long term of doing an experiment over and over, you would **expect** this average.

You toss a coin and record the result. What is the probability that the result is heads? If you flip a coin two times, does probability tell you that these flips will result in one heads and one tail? You might toss a fair coin ten times and record nine heads. As you learned in Chapter 3, probability does not describe the short-term results of an experiment. It gives information about what can be expected in the long term. To demonstrate this, Karl Pearson once tossed a fair coin 24,000 times! He recorded the results of each toss, obtaining heads 12,012 times. **In his experiment, Pearson illustrated the Law of Large Numbers**.

The Law of Large Numbers states that, as the number of trials in a probability experiment increases, the difference between the theoretical probability of an event and the relative frequency approaches zero **(the theoretical probability and the relative frequency get closer and closer together)**. When evaluating the long-term results of statistical experiments, we often want to know the “average” outcome. This “long-term average” is known as the **mean** or **expected value** of the experiment and is denoted by the Greek letter μ. In other words, after conducting many trials of an experiment, you would expect this average value.

Note:
NOTE
To find the expected value or long term average, μ, simply multiply each value of the random variable by its probability and add the products.

Generally for probability distributions, we use a calculator or a computer to calculate μ and σ to reduce roundoff error. For some probability distributions, there are short-cut formulas for calculating μ and σ.

Example:
A men's soccer team plays soccer zero, one, or two days a week. The probability that they play zero days is 0.2, the probability that they play one day is 0.5, and the probability that they play two days is 0.3. Find the long-term average or expected value, μ, of the number of days per week the men's soccer team plays soccer.

To do the problem, first let the random variable X = the number of days the men's soccer team plays soccer per week. Then x takes on the values 0, 1, 2. Construct a PDF table adding a column $x*P(x)$. In this column, you will multiply each x value by its probability.

x	P**(*x*)**	***x****P**(*x*)**
0	0.2	(0)(0.2) = 0
1	0.5	(1)(0.5) = 0.5
2	0.3	(2)(0.3) = 0.6

Expected Value TableThis table is called an expected value table. The table helps you calculate the expected value or long-term average.

Add the last column $x*P(x)$ to find the long term average or expected value: (0)(0.2) + (1)(0.5) + (2)(0.3) = 0 + 0.5 + 0.6 = 1.1.

The expected value is 1.1. The men's soccer team would, on the average, expect to play soccer 1.1 days per week. The number 1.1 is the long-term average or expected value if the men's soccer team plays soccer week after week after week. We say μ = 1.1.

Like data, probability distributions have standard deviations. To calculate the standard deviation (σ) of a probability distribution, find each deviation from its expected value, square it, multiply it by its probability, add the products, and take the square root. To understand how to do the calculation, look at the table for the number of days per week a men's soccer team plays soccer. To find the standard deviation, add the entries in the column labeled $(x – \mu)^2P(x)$ and take the square root.

x	P**(*x*)**	***x****P**(*x*)**	**(*x* – *μ*)^2**P**(*x*)**
0	0.2	(0)(0.2) = 0	$(0 – 1.1)^2(0.2) = 0.242$
1	0.5	(1)(0.5) = 0.5	$(1 – 1.1)^2(0.5) = 0.005$
2	0.3	(2)(0.3) = 0.6	$(2 – 1.1)^2(0.3) = 0.243$

Add the last column in the table. 0.242 + 0.005 + 0.243 = 0.490. The standard deviation is the square root of 0.49, or $\sigma = \sqrt{0.49} = 0.7$

Example:

Find the expected value of the number of times a newborn baby's crying wakes its mother after midnight. The expected value is the expected number of times per week a newborn baby's crying wakes its mother after midnight. Calculate the standard deviation of the variable as well.

x	**$P(x)$**	**$x*P(x)$**	**$(x-\mu)^2 \cdot P(x)$**
0	$P(x=0) = \frac{2}{50}$	$(0)\left(\frac{2}{50}\right) = 0$	$(0-2.1)^2 \cdot 0.04 = 0.1764$
1	$P(x=1) = \left(\frac{11}{50}\right)$	$(1)\left(\frac{11}{50}\right) = \frac{11}{50}$	$(1-2.1)^2 \cdot 0.22 = 0.2662$
2	$P(x=2) = \frac{23}{50}$	$(2)\left(\frac{23}{50}\right) = \frac{46}{50}$	$(2-2.1)^2 \cdot 0.46 = 0.0046$
3	$P(x=3) = \frac{9}{50}$	$(3)\left(\frac{9}{50}\right) = \frac{27}{50}$	$(3-2.1)^2 \cdot 0.18 = 0.1458$
4	$P(x=4) = \frac{4}{50}$	$(4)\left(\frac{4}{50}\right) = \frac{16}{50}$	$(4-2.1)^2 \cdot 0.08 = 0.2888$
5	$P(x=5) = \frac{1}{50}$	$(5)\left(\frac{1}{50}\right) = \frac{5}{50}$	$(5-2.1)^2 \cdot 0.02 = 0.1682$

You expect a newborn to wake its mother after midnight 2.1 times per week, on the average.

Add the values in the third column of the table to find the expected value of X:

μ = Expected Value = $\frac{105}{50} = 2.1$

Use μ to complete the table. The fourth column of this table will provide the values you need to calculate the standard deviation. For each value x, multiply the square of its deviation by its probability. (Each deviation has the format $x - \mu$).

Add the values in the fourth column of the table:

0.1764 + 0.2662 + 0.0046 + 0.1458 + 0.2888 + 0.1682 = 1.05

The standard deviation of x is the square root of this sum: $\sigma = \sqrt{1.05} \approx 1.0247$

Note:
Try It
Exercise:

Problem:

A hospital researcher is interested in the number of times the average post-op patient will ring the nurse during a 12-hour shift. For a random sample of 50 patients, the following information was obtained. What is the expected value?

x	***P*(*x*)**
0	$P(x = 0) = \frac{4}{50}$
1	$P(x = 1) = \frac{8}{50}$
2	$P(x = 2) = \frac{16}{50}$
3	$P(x = 3) = \frac{14}{50}$
4	$P(x = 4) = \frac{6}{50}$

x	***P*(*x*)**
5	$P(x = 5) = \frac{2}{50}$

Solution:

The expected value is 2.24

$(0)\frac{4}{50} + (1)\frac{8}{50} + (2)\frac{16}{50} + (3)\frac{14}{50} + (4)\frac{6}{50} + (5)\frac{2}{50} = 0 + \frac{8}{50} + \frac{32}{50} + \frac{42}{50} + \frac{24}{50} + \frac{10}{50} = \frac{116}{50} = 2.32$

Example:
Suppose you play a game of chance in which five numbers are chosen from 0, 1, 2, 3, 4, 5, 6, 7, 8, 9. A computer randomly selects five numbers from zero to nine with replacement. You pay \$2 to play and could profit \$100,000 if you match all five numbers in order (you get your \$2 back plus \$100,000). Over the long term, what is your **expected** profit of playing the game?

To do this problem, set up an expected value table for the amount of money you can profit.

Let X = the amount of money you profit. The values of x are not 0, 1, 2, 3, 4, 5, 6, 7, 8, 9. Since you are interested in your profit (or loss), the values of x are 100,000 dollars and −2 dollars.

To win, you must get all five numbers correct, in order. The probability of choosing one correct number is $\frac{1}{10}$ because there are ten numbers. You may choose a number more than once. The probability of choosing all five numbers correctly and in order is
Equation:

$$\left(\frac{1}{10}\right)\left(\frac{1}{10}\right)\left(\frac{1}{10}\right)\left(\frac{1}{10}\right)\left(\frac{1}{10}\right) = (1)(10^{-5}) = 0.00001.$$

Therefore, the probability of winning is 0.00001 and the probability of losing is

Equation:

$$1 - 0.00001 = 0.99999.$$

The expected value table is as follows:

	x	P**(*x*)**	***x****P**(*x*)**
Loss	–2	0.99999	(–2)(0.99999) = –1.99998
Profit	100,000	0.00001	(100000)(0.00001) = 1

Add the last column. –1.99998 + 1 = –0.99998

Since –0.99998 is about –1, you would, on average, expect to lose approximately $1 for each game you play. However, each time you play, you either lose $2 or profit $100,000. The $1 is the average or expected LOSS per game after playing this game over and over.

Note:

Try It

Exercise:

Problem:

You are playing a game of chance in which four cards are drawn from a standard deck of 52 cards. You guess the suit of each card before it is drawn. The cards are replaced in the deck on each draw. You pay $1 to play. If you guess the right suit every time, you get your money back and $256. What is your expected profit of playing the game over the long term?

Solution:

Let X = the amount of money you profit. The x-values are –$1 and $256.

The probability of guessing the right suit each time is $\left(\frac{1}{4}\right)\left(\frac{1}{4}\right)\left(\frac{1}{4}\right)\left(\frac{1}{4}\right)=\frac{1}{256}$ = 0.0039

The probability of losing is 1 – $\frac{1}{256}$ = $\frac{255}{256}$ = 0.9961

(0.0039)256 + (0.9961)(–1) = 0.9984 + (–0.9961) = 0.0023 or 0.23 cents.

Example:

Suppose you play a game with a biased coin. You play each game by tossing the coin once. $P(\text{heads})$ = $\frac{2}{3}$ and $P(\text{tails})$ = $\frac{1}{3}$. If you toss a head, you pay $6. If you toss a tail, you win $10. If you play this game many times, will you come out ahead?

Exercise:

Problem: a. Define a random variable X.

Solution:

a. X = amount of profit

Exercise:

Problem: b. Complete the following expected value table.

	x	____	____
WIN	10	$\frac{1}{3}$	____
LOSE	____	____	$\frac{-12}{3}$

Solution:

b.

	x	P**(*x*)**	***x****P**(*x*)**
WIN	10	$\frac{1}{3}$	$\frac{10}{3}$
LOSE	–6	$\frac{2}{3}$	$\frac{-12}{3}$

Exercise:

Problem: c. What is the expected value, μ? Do you come out ahead?

Solution:

c. Add the last column of the table. The expected value $\mu = \frac{-2}{3}$. You lose, on average, about 67 cents each time you play the game so you do not come out ahead.

Note:
Try It
Exercise:

Problem:

Suppose you play a game with a spinner. You play each game by spinning the spinner once. $P(\text{red}) = \frac{2}{5}$, $P(\text{blue}) = \frac{2}{5}$, and $P(\text{green}) = \frac{1}{5}$. If you land on red, you pay $10. If you land on blue, you don't pay or win anything. If you land on green, you win $10. Complete the following expected value table.

	x	P**(*x*)**	
Red			$-\frac{20}{5}$
Blue		$\frac{2}{5}$	
Green	10		

Solution:

	x	***P*(*x*)**	***x***P*(*x*)**
Red	–10	$\frac{2}{5}$	$-\frac{20}{5}$
Blue	0	$\frac{2}{5}$	$\frac{0}{5}$
Green	10	$\frac{1}{5}$	$\frac{10}{5}$

Example:
Exercise:

Problem:

Toss a fair, six-sided die twice. Let X = the number of faces that show an even number. Construct a table like [link] and calculate the mean μ and standard deviation σ of X.

Solution:

Tossing one fair six-sided die twice has the same sample space as tossing two fair six-sided dice. The sample space has 36 outcomes:

(1, 1)	(1, 2)	(1, 3)	(1, 4)	(1, 5)	(1, 6)
(2, 1)	(2, 2)	(2, 3)	(2, 4)	(2, 5)	(2, 6)
(3, 1)	(3, 2)	(3, 3)	(3, 4)	(3, 5)	(3, 6)
(4, 1)	(4, 2)	(4, 3)	(4, 4)	(4, 5)	(4, 6)
(5, 1)	(5, 2)	(5, 3)	(5, 4)	(5, 5)	(5, 6)
(6, 1)	(6, 2)	(6, 3)	(6, 4)	(6, 5)	(6, 6)

Use the sample space to complete the following table:

x	**$P(x)$**	**$x*P(x)$**	**$(x-\mu)^2 \cdot P(x)$**
0	$\frac{9}{36}$	0	$(0-1)^2 \cdot \frac{9}{36} = \frac{9}{36}$
1	$\frac{18}{36}$	$\frac{18}{36}$	$(1-1)^2 \cdot \frac{18}{36} = 0$
2	$\frac{9}{36}$	$\frac{18}{36}$	$(2-1)^2 \cdot \frac{9}{36} = \frac{9}{36}$

Calculating μ and σ.

Add the values in the third column to find the expected value: $\mu = \frac{36}{36} = 1$. Use this value to complete the fourth column.

Add the values in the fourth column and take the square root of the sum: $\sigma = \sqrt{\frac{18}{36}} \approx 0.7071$.

Example:
Exercise:

Problem:

On May 11, 2013 at 9:30 PM, the probability that moderate seismic activity (one moderate earthquake) would occur in the next 48 hours in Iran was about 21.42%. Suppose you make a bet that a moderate earthquake will occur in Iran during this period. If you win the bet, you win \$50. If you lose the bet, you pay \$20. Let X = the amount of profit from a bet.

$P(\text{win}) = P(\text{one moderate earthquake will occur}) = 21.42\%$

$P(\text{loss}) = P(\text{one moderate earthquake will NOT occur}) = 100\% – 21.42\%$

If you bet many times, will you come out ahead? Explain your answer in a complete sentence using numbers. What is the standard deviation of X? Construct a table similar to [link] and [link] to help you answer these questions.

Solution:

	x	$P(x)$	$x*P(x)$	$(x – \mu)^2P(x)$
win	50	0.2142	10.71	$[50 – (–5.006)]^2(0.2142) = 648.0964$

	x	P**(*x*)**	***x****P**(*x*)**	**(*x* – *μ*)^2**P**(*x*)**
loss	–20	0.7858	–15.716	[–20 – (–5.006)]^2(0.7858) = 176.6636

Mean = Expected Value = 10.71 + (–15.716) = –5.006.

If you make this bet many times under the same conditions, your long term outcome will be an average *loss* of $5.01 per bet.

$\text{Standard Deviation} = \sqrt{648.0964 + 176.6636} \approx 28.7186$

Note:
Try It
Exercise:

Problem:

On May 11, 2013 at 9:30 PM, the probability that moderate seismic activity (one moderate earthquake) would occur in the next 48 hours in Japan was about 1.08%. As in [link], you bet that a moderate earthquake will occur in Japan during this period. If you win the bet, you win $100. If you lose the bet, you pay $10. Let X = the amount of profit from a bet. Find the mean and standard deviation of X.

Solution:

	x	$P(x)$	$x*P(x)$	$(x - \mu)^2P(x)$
win	100	0.0108	1.08	$[100 - (-8.812)]^2 \cdot 0.0108 = 127.8726$
loss	–10	0.9892	–9.892	$[-10 - (-8.812)]^2 \cdot 0.9892 = 1.3961$

Mean = Expected Value = μ = 1.08 + (–9.892) = –8.812

If you make this bet many times under the same conditions, your long term outcome will be an average loss of $8.81 per bet.

$$\text{Standard Deviation} = \sqrt{127.7826 + 1.3961} \approx 11.3696$$

Some of the more common discrete probability functions are binomial, geometric, hypergeometric, and Poisson. Most elementary courses do not cover the geometric, hypergeometric, and Poisson. Your instructor will let you know if he or she wishes to cover these distributions.

A probability distribution function is a pattern. You try to fit a probability problem into a **pattern** or distribution in order to perform the necessary calculations. These distributions are tools to make solving probability problems easier. Each distribution has its own special characteristics. Learning the characteristics enables you to distinguish among the different distributions.

References

Class Catalogue at the Florida State University. Available online at https://apps.oti.fsu.edu/RegistrarCourseLookup/SearchFormLegacy (accessed May 15, 2013).

“World Earthquakes: Live Earthquake News and Highlights,” World Earthquakes, 2012. http://www.world-earthquakes.com/index.php?option=ethq_prediction (accessed May 15, 2013).

Section Review

The expected value, or mean, of a discrete random variable predicts the long-term results of a statistical experiment that has been repeated many times. The standard deviation of a probability distribution is used to measure the variability of possible outcomes.

Formula Review

Mean or Expected Value: $\mu = \sum_{x \in X} xP(x)$

Standard Deviation: $\sigma = \sqrt{\sum_{x \in X} (x-\mu)^2 P(x)}$

Exercise:

Problem: Complete the expected value table.

x	P**(*x*)**	***x****P**(*x*)**
0	0.2	
1	0.2	
2	0.4	
3	0.2	

Exercise:

Problem: Find the expected value from the expected value table.

x	$P(x)$	$x*P(x)$
2	0.1	2(0.1) = 0.2
4	0.3	4(0.3) = 1.2
6	0.4	6(0.4) = 2.4
8	0.2	8(0.2) = 1.6

Solution:

0.2 + 1.2 + 2.4 + 1.6 = 5.4

Exercise:

Problem: Find the standard deviation.

x	$P(x)$	$x*P(x)$	$(x-\mu)^2P(x)$

x	$P(x)$	$x*P(x)$	$(x - \mu)^2P(x)$
2	0.1	(2)(0.1) = 0.2	$(2 - 5.4)^2(0.1) = 1.156$
4	0.3	(4)(0.3) = 1.2	$(4 - 5.4)^2(0.3) = 0.588$
6	0.4	(6)(0.4) = 2.4	$(6 - 5.4)^2(0.4) = 0.144$
8	0.2	(8)(0.2) = 1.6	$(8 - 5.4)^2(0.2) = 1.352$

Exercise:

Problem: Identify the mistake in the probability distribution table.

x	$P(x)$	$x*P(x)$
1	0.15	0.15
2	0.25	0.50
3	0.30	0.90
4	0.20	0.80
5	0.15	0.75

Solution:

The values of $P(x)$ do not sum to one.

Exercise:

Problem: Identify the mistake in the probability distribution table.

x	***P*(*x*)**	***x*****P*(*x*)**
1	0.15	0.15
2	0.25	0.40
3	0.25	0.65
4	0.20	0.85
5	0.15	1

Use the following information to answer the next five exercises: A physics professor wants to know what percent of physics majors will spend the next several years doing post-graduate research. He has the following probability distribution.

x	***P*(*x*)**	***x*****P*(*x*)**
1	0.35	

x	P(*x*)	*x**P(*x*)
2	0.20	
3	0.15	
4		
5	0.10	
6	0.05	

Exercise:

Problem: Define the random variable X.

Solution:

Let X = the number of years a physics major will spend doing post-graduate research.

Exercise:

Problem: Define $P(x)$, or the probability of x.

Exercise:

Problem:

Find the probability that a physics major will do post-graduate research for four years. $P(x = 4)$ = _______

Solution:

1 – 0.35 – 0.20 – 0.15 – 0.10 – 0.05 = 0.15

Exercise:

Problem:

FInd the probability that a physics major will do post-graduate research for at most three years. $P(x \leq 3)$ = _______

Exercise:

Problem:

On average, how many years would you expect a physics major to spend doing post-graduate research?

Solution:

1(0.35) + 2(0.20) + 3(0.15) + 4(0.15) + 5(0.10) + 6(0.05) = 0.35 + 0.40 + 0.45 + 0.60 + 0.50 + 0.30 = 2.6 years

Use the following information to answer the next seven exercises: A ballet instructor is interested in knowing what percent of each year's class will continue on to the next, so that she can plan what classes to offer. Over the years, she has established the following probability distribution.

- Let X = the number of years a student will study ballet with the teacher.
- Let $P(x)$ = the probability that a student will study ballet x years.

Exercise:

Problem: Complete the following table using the data provided.

x	P(*x*)	*x**P(*x*)

x	P**(*x*)**	***x****P**(*x*)**
1	0.10	
2	0.05	
3	0.10	
4		
5	0.30	
6	0.20	
7	0.10	

Exercise:

Problem: In words, define the random variable X.

Solution:

X is the number of years a student studies ballet with the teacher.

Exercise:

Problem: $P(x = 4)$ = _______

Exercise:

Problem: $P(x < 4)$ = _______

Solution:

0.10 + 0.05 + 0.10 = 0.25

Exercise:

Problem:

On average, how many years would you expect a child to study ballet with this teacher?

Exercise:

Problem: What does the column "$P(x)$" sum to and why?

Solution:

The sum of the probabilities sum to one because it is a probability distribution.

Exercise:

Problem: What does the column "$x*P(x)$" sum to and why?

Exercise:

Problem:

You are playing a game by drawing a card from a standard deck and replacing it. If the card is a face card, you win $30. If it is not a face card, you pay $2. There are 12 face cards in a deck of 52 cards. What is the expected value of playing the game?

Solution:

$$-2\left(\frac{40}{52}\right)+30\left(\frac{12}{52}\right)=-1.54+6.92=5.38$$

Exercise:

Problem:

You are playing a game by drawing a card from a standard deck and replacing it. If the card is a face card, you win $30. If it is not a face card, you pay $2. There are 12 face cards in a deck of 52 cards. Should you play the game?

HOMEWORK

Exercise:

Problem:

A theater group holds a fund-raiser. It sells 100 raffle tickets for $5 apiece. Suppose you purchase four tickets. The prize is two passes to a Broadway show, worth a total of $150.

a. What are you interested in here?
b. In words, define the random variable X.
c. List the values that X may take on.
d. Construct a PDF.
e. If this fund-raiser is repeated often and you always purchase four tickets, what would be your expected average winnings per raffle?

Exercise:

Problem:

A game involves selecting a card from a regular 52-card deck and tossing a coin. The coin is a fair coin and is equally likely to land on heads or tails.

- If the card is a face card, and the coin lands on Heads, you win $6
- If the card is a face card, and the coin lands on Tails, you win $2
- If the card is not a face card, you lose $2, no matter what the coin shows.

a. Find the expected value for this game (expected net gain or loss).
b. Explain what your calculations indicate about your long-term average profits and losses on this game.
c. Should you play this game to win money?

Solution:

The variable of interest is X, or the gain or loss, in dollars.

The face cards jack, queen, and king. There are (3)(4) = 12 face cards and 52 – 12 = 40 cards that are not face cards.

We first need to construct the probability distribution for X. We use the card and coin events to determine the probability for each outcome, but we use the monetary value of X to determine the expected value.

Card Event	***X* net gain/loss**	P**(*X*)**
Face Card and Heads	6	$\left(\frac{12}{52}\right)\left(\frac{1}{2}\right)=\left(\frac{6}{52}\right)$
Face Card and Tails	2	$\left(\frac{12}{52}\right)\left(\frac{1}{2}\right)=\left(\frac{6}{52}\right)$
(Not Face Card) and (H or T)	–2	$\left(\frac{40}{52}\right)(1)=\left(\frac{40}{52}\right)$

- $\text{Expected value}=(6)\left(\frac{6}{52}\right)+(2)\left(\frac{6}{52}\right)+(-2)\left(\frac{40}{52}\right)=-\frac{32}{52}$
- Expected value = –$0.62, rounded to the nearest cent
- If you play this game repeatedly, over a long string of games, you would expect to lose 62 cents per game, on average.
- You should not play this game to win money because the expected value indicates an expected average loss.

Exercise:

Problem:

You buy a lottery ticket to a lottery that costs $10 per ticket. There are only 100 tickets available to be sold in this lottery. In this lottery there are one $500 prize, two $100 prizes, and four $25 prizes. Find your expected gain or loss.

Exercise:

Problem: Complete the PDF and answer the questions.

x	$P(x)$	$x*P(x)$
0	0.3	
1	0.2	
2		
3	0.4	

a. Find the probability that $x = 2$.
b. Find the expected value.

Solution:

a. 0.1
b. 1.6

Exercise:

Problem:

Suppose that you are offered the following “deal.” You roll a die. If you roll a six, you win $10. If you roll a four or five, you win $5. If you roll a one, two, or three, you pay $6.

a. What are you ultimately interested in here (the value of the roll or the money you win)?
b. In words, define the Random Variable X.
c. List the values that x may take on.
d. Construct a PDF.
e. Over the long run of playing this game, what are your expected average winnings per game?
f. Based on numerical values, should you take the deal? Explain your decision in complete sentences.

Exercise:

Problem:

A venture capitalist, willing to invest $1,000,000, has three investments to choose from. The first investment, a software company, has a 10% chance of returning $5,000,000 profit, a 30% chance of returning $1,000,000 profit, and a 60% chance of losing the million dollars. The second company, a hardware company, has a 20% chance of returning $3,000,000 profit, a 40% chance of returning $1,000,000 profit, and a 40% chance of losing the million dollars. The third company, a biotech firm, has a 10% chance of returning $6,000,000 profit, a 70% of no profit or loss, and a 20% chance of losing the million dollars.

a. Construct a PDF for each investment.
b. Find the expected value for each investment.
c. Which is the safest investment? Why do you think so?
d. Which is the riskiest investment? Why do you think so?
e. Which investment has the highest expected return, on average?

Solution:

a.

Software Company	
x	$P(x)$
5,000,000	0.10
1,000,000	0.30
–1,000,000	0.60

Hardware Company	
x	$P(x)$
3,000,000	0.20
1,000,000	0.40
–1,000,00	0.40

Biotech Firm	
x	$P(x)$
6,00,000	0.10

Biotech Firm	
x	***P*(*x*)**
0	0.70
–1,000,000	0.20

b. $200,000; $600,000; $400,000
c. third investment because it has the lowest probability of loss
d. first investment because it has the highest probability of loss
e. second investment

Exercise:

Problem:

Suppose that 20,000 married adults in the United States were randomly surveyed as to the number of children they have. The results are compiled and are used as theoretical probabilities. Let X = the number of children married people have.

x	***P*(*x*)**	***x**P*(*x*)**
0	0.10	
1	0.20	
2	0.30	
3		

x	***P*(*x*)**	***x**P*(*x*)**
4	0.10	
5	0.05	
6 (or more)	0.05	

a. Find the probability that a married adult has three children.
b. In words, what does the expected value in this example represent?
c. Find the expected value.
d. Is it more likely that a married adult will have two to three children or four to six children? How do you know?

Exercise:

Problem:

Suppose that the PDF for the number of years it takes to earn a Bachelor of Science (B.S.) degree is given as in the following table.

x	***P*(*x*)**
3	0.05
4	0.40
5	0.30
6	0.15

x	***P*(*x*)**
7	0.10

On average, how many years do you expect it to take for an individual to earn a B.S.?

Solution:

4.85 years

Exercise:

Problem:

People visiting video rental stores often rent more than one DVD at a time. The probability distribution for DVD rentals per customer at Video To Go is given in the following table. There is a five-video limit per customer at this store, so nobody ever rents more than five DVDs.

x	***P*(*x*)**
0	0.03
1	0.50
2	0.24
3	
4	0.07

x	$P(x)$
5	0.04

a. Describe the random variable X in words.
b. Find the probability that a customer rents three DVDs.
c. Find the probability that a customer rents at least four DVDs.
d. Find the probability that a customer rents at most two DVDs. Another shop, Entertainment Headquarters, rents DVDs and video games. The probability distribution for DVD rentals per customer at this shop is given as follows. They also have a five-DVD limit per customer.

x	$P(x)$
0	0.35
1	0.25
2	0.20
3	0.10
4	0.05
5	0.05

e. At which store is the expected number of DVDs rented per customer higher?
f. If Video to Go estimates that they will have 300 customers next week, how many DVDs do they expect to rent next week?

Answer in sentence form.

g. If Video to Go expects 300 customers next week, and Entertainment HQ projects that they will have 420 customers, for which store is the expected number of DVD rentals for next week higher? Explain.
h. Which of the two video stores experiences more variation in the number of DVD rentals per customer? How do you know that?

Exercise:

Problem:

A “friend” offers you the following “deal.” For a $10 fee, you may pick an envelope from a box containing 100 seemingly identical envelopes. However, each envelope contains a coupon for a free gift.

- Ten of the coupons are for a free gift worth $6.
- Eighty of the coupons are for a free gift worth $8.
- Six of the coupons are for a free gift worth $12.
- Four of the coupons are for a free gift worth $40.

Based upon the financial gain or loss over the long run, should you play the game?

a. Yes, I expect to come out ahead in money.
b. No, I expect to come out behind in money.
c. It doesn’t matter. I expect to break even.

Solution:

b

Exercise:

Problem:

Florida State University has 14 statistics classes scheduled for its Summer 2013 term. One class has space available for 30 students, eight classes have space for 60 students, one class has space for 70 students, and four classes have space for 100 students.

a. What is the average class size assuming each class is filled to capacity?
b. Space is available for 980 students. Suppose that each class is filled to capacity and select a statistics student at random. Let the random variable X equal the size of the student's class. Define the PDF for X.
c. Find the mean of X.
d. Find the standard deviation of X.

Exercise:

Problem:

In a lottery, there are 250 prizes of \$5, 50 prizes of \$25, and ten prizes of \$100. Assuming that 10,000 tickets are to be issued and sold, what is a fair price to charge to break even?

Solution:

Let X = the amount of money to be won on a ticket. The following table shows the PDF for X.

x	P**(*x*)**
0	0.969

x	P**(*x*)**
5	$\frac{250}{10{,}000}$ = 0.025
25	$\frac{50}{10{,}000}$ = 0.005
100	$\frac{10}{10{,}000}$ = 0.001

Calculate the expected value of X.

0(0.969) + 5(0.025) + 25(0.005) + 100(0.001) = 0.35

A fair price for a ticket is $0.35. Any price over $0.35 will enable the lottery to raise money.

Glossary

Expected Value

expected arithmetic average when an experiment is repeated many times; also called the mean. Notations: μ. For a discrete random variable (RV) with probability distribution function $P(x)$,the definition can also be written in the form $\mu = \sum xP\left(x\right)$.

Mean

a number that measures the central tendency; a common name for mean is ‘average.’ The term ‘mean’ is a shortened form of ‘arithmetic mean.’ By definition, the mean for a sample (denoted by x) is $x = \frac{\text{Sum of all values in the sample}}{\text{Number of values in the sample}}$ and the mean for a population (denoted by μ) is $\mu = \frac{\text{Sum of all values in the population}}{\text{Number of values in the population}}$.

Mean of a Probability Distribution

the long-term average of many trials of a statistical experiment

Standard Deviation of a Probability Distribution

a number that measures how far the outcomes of a statistical experiment are from the mean of the distribution

The Law of Large Numbers

As the number of trials in a probability experiment increases, the difference between the theoretical probability of an event and the relative frequency probability approaches zero.

The Binomial Distribution

There are three characteristics of a binomial experiment.

1. There are a fixed number of trials. Think of trials as repetitions of an experiment. The letter n denotes the number of trials.
2. There are only two possible outcomes, called "success" and "failure," for each trial. The letter p denotes the probability of a success on one trial, and q denotes the probability of a failure on one trial. $p + q = 1$.
3. The n trials are independent and are repeated using identical conditions. Because the n trials are independent, the outcome of one trial does not help in predicting the outcome of another trial. Another way of saying this is that for each individual trial, the probability, p, of a success and probability, q, of a failure remain the same. For example, randomly guessing at a true-false statistics question has only two outcomes. If a success is guessing correctly, then a failure is guessing incorrectly. Suppose Joe always guesses correctly on any statistics true-false question with probability $p = 0.6$. Then, $q = 0.4$. This means that for every true-false statistics question Joe answers, his probability of success ($p = 0.6$) and his probability of failure ($q = 0.4$) remain the same.

The outcomes of a binomial experiment fit a **binomial probability distribution**. The random variable X = the number of successes obtained in the n independent trials.

The mean, μ, and variance, σ^2, for the binomial probability distribution are $\mu = np$ and $\sigma^2 = npq$. The standard deviation, σ, is then $\sigma = \sqrt{npq}$.

Any experiment that has characteristics two and three and where $n = 1$ is called a **Bernoulli Trial** (named after Jacob Bernoulli who, in the late 1600s, studied them extensively). A binomial experiment takes place when the number of successes is counted in one or more Bernoulli Trials.

Example:
At ABC College, the withdrawal rate from an elementary physics course is 30% for any given term. This implies that, for any given term, 70% of the students stay in the class for the entire term. A "success" could be defined as an individual who withdrew. The random variable X = the number of students who withdraw from the randomly selected elementary physics class.

Note:
Try It
Exercise:

Problem:

The state health board is concerned about the amount of fruit available in school lunches. Forty-eight percent of schools in the state offer fruit in their lunches every day. This implies that 52% do not. What would a "success" be in this case?

Solution:

a school that offers fruit in their lunch every day

Example:
Suppose you play a game that you can only either win or lose. The probability that you win any game is 55%, and the probability that you lose is 45%. Each game you play is independent. If you play the game 20 times, write the function that describes the probability that you win 15 of the 20 times. Here, if you define X as the number of wins, then X takes on the values 0, 1, 2, 3, ..., 20. The probability of a success is p = 0.55. The probability of a failure is q = 0.45. The number of trials is n = 20. The probability question can be stated mathematically as $P(x = 15)$.

Note:
Try It
Exercise:

Problem:

A trainer is teaching a dolphin to do tricks. The probability that the dolphin successfully performs the trick is 35%, and the probability that the dolphin does not successfully perform the trick is 65%. Out of 20 attempts, you want to find the probability that the dolphin succeeds 12 times. State the probability question mathematically.

Solution:

$P(x = 12)$

Example:
Exercise:

Problem:

A fair coin is flipped 15 times. Each flip is independent. What is the probability of getting more than ten heads? Let X = the number of heads in 15 flips of the fair coin. X takes on the values 0, 1, 2, 3, ..., 15. Since the coin is fair, p = 0.5 and q = 0.5. The number of trials is n = 15. State the probability question mathematically.

Solution:

$P(x > 10)$

Note:
Try It
Exercise:

Problem:

A fair, six-sided die is rolled ten times. Each roll is independent. You want to find the probability of rolling a one more than three times. State the probability question mathematically.

Solution:

$P(x > 3)$

Example:

Approximately 70% of statistics students do their homework in time for it to be collected and graded. Each student does homework independently. In a statistics class of 50 students, what is the probability that at least 40 will do their homework on time? Students are selected randomly.

Exercise:

Problem:

a. This is a binomial problem because there is only a success or a __________, there are a fixed number of trials, and the probability of a success is 0.70 for each trial.

Solution:

a. failure

Exercise:

Problem:

b. If we are interested in the number of students who do their homework on time, then how do we define X?

Solution:

b. X = the number of statistics students who do their homework on time

Exercise:

Problem: c. What values does x take on?

Solution:

c. 0, 1, 2, …, 50

Exercise:

Problem: d. What is a "failure," in words?

Solution:

d. Failure is defined as a student who does not complete his or her homework on time.

The probability of a success is p = 0.70. The number of trials is n = 50.

Exercise:

Problem: e. If p + q = 1, then what is q?

Solution:

e. q = 0.30

Exercise:

Problem:

f. The words "at least" translate as what kind of inequality for the probability question
$P(x \_\_\_\_ 40)$.

Solution:

f. greater than or equal to (≥)
The probability question is $P(x \geq 40)$.

Note:
Try It
Exercise:

Problem:

Sixty-five percent of people pass the state driver’s exam on the first try. A group of 50 individuals who have taken the driver’s exam is randomly selected. Give two reasons why this is a binomial problem.

Solution:

This is a binomial problem because there is only a success or a failure, and there are a definite number of trials. The probability of a success stays the same for each trial.

Notation for the Binomial: *B* = Binomial Probability Distribution Function

$X \sim B(n, p)$

Read this as "X is a random variable with a binomial distribution." The parameters are n and p; n = number of trials, p = probability of a success on each trial.

The mean of the binomial distribution is $\mu = np$.

The standard deviation of the binomial distribution is $\sigma = \sqrt{npq}$.

Example:
It has been stated that about 41% of adult workers have a high school diploma but do not pursue any further education. If 20 adult workers are randomly selected, find the probability that at most 12 of them have a high school diploma but do not pursue any further education. How many adult workers do you expect to have a high school diploma but do not pursue any further education?

Let X = the number of workers who have a high school diploma but do not pursue any further education.

X takes on the values 0, 1, 2, ..., 20 where $n = 20$, $p = 0.41$, and $q = 1 – 0.41 = 0.59$. $X \sim B(20, 0.41)$

Find $P(x \leq 12)$. $P(x \leq 12) = 0.9738$. (See instructions below on how to find this probability using the calculator.)

Note:
On the TI-83/84 calculator, go into 2nd DISTR and scroll down to A and B to find the "binompdf" and "binomcdf" functions. The instructions for using these functions are as follows:
To calculate $P(x = \text{number})$: binompdf(*n*, *p*, number) if "number" is left out, the result is the binomial probability table.

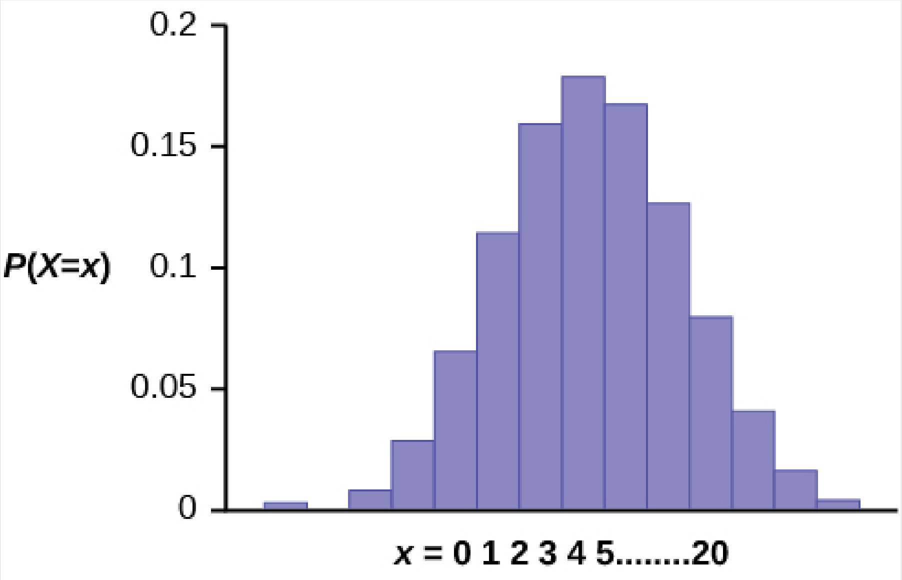
To calculate $P(x \leq \text{number})$: binomcdf(*n*, *p*, number) if "number" is left out, the result is the cumulative binomial probability table.

For the problem above: After you are in 2nd DISTR, scroll down to binomcdf. Press ENTER. Enter binomcdf(20,0.41,12).The result is $P(x \leq 12) = 0.9738$.

Note:
NOTE
If you want to find $P(x = 12)$, use the pdf (binompdf). If you want to find $P(x > 12)$, use 1 - binomcdf(20,0.41,12).

The probability that at most 12 workers have a high school diploma but do not pursue any further education is 0.9738.

The graph of $X \sim B(20, 0.41)$ is as follows:



The y-axis contains the probability of x, where X = the number of workers who have only a high school diploma.

The number of adult workers that you expect to have a high school diploma but not pursue any further education is the mean, $\mu = np$ = (20)(0.41) = 8.2 workers.

The formula for the variance is $\sigma^2 = npq$. The standard deviation is $\sigma = \sqrt{npq}$.
$\sigma = \sqrt{(20)(0.41)(0.59)}$ = 2.20 workers.

Note:
Try It
Exercise:

Problem:

About 32% of students participate in a community volunteer program outside of school. If 30 students are selected at random, find the probability that at most 14 of them participate in a community volunteer program outside of school. Use the TI-83+ or TI-84 calculator to find the answer.

Solution:

$P(x \le 14)$ = 0.9695

Example:
Exercise:

Problem:

In the 2013 *Jerry's Artarama* art supplies catalog, there are 560 pages. Eight of the pages feature signature artists. Suppose we randomly sample 100 pages. Let X = the number of pages that feature signature artists.

a. What values does x take on?
b. What is the probability distribution? Find the following probabilities:

    i. the probability that two pages feature signature artists
    ii. the probability that at most six pages feature signature artists
    iii. the probability that more than three pages feature signature artists.

c. Using the formulas, calculate the (i) mean and (ii) standard deviation.

Solution:

a. x = 0, 1, 2, 3, 4, 5, 6, 7, 8
b. $X \sim B\left(100, \frac{8}{560}\right)$

    i. $P(x = 2)$ = binompdf$\left(100, \frac{8}{560}, 2\right)$ = 0.2466
    ii. $P(x \leq 6)$ = binomcdf$\left(100, \frac{8}{560}, 6\right)$ = 0.9994
    iii. $P(x > 3)$ = 1 – $P(x \leq 3)$ = 1 – binomcdf$\left(100, \frac{8}{560}, 3\right)$ = 1 – 0.9443 = 0.0557

c.
    i. Mean = np = $(100)\left(\frac{8}{560}\right) = \frac{800}{560}$ ≈ 1.4286
    ii. Standard Deviation = $\sqrt{npq} = \sqrt{(100)\left(\frac{8}{560}\right)\left(\frac{552}{560}\right)}$ ≈ 1.1867

Note:
Try It
Exercise:

Problem:

According to a Gallup poll, 60% of American adults prefer saving over spending. Let X = the number of American adults out of a random sample of 50 who prefer saving to spending.

a. What is the probability distribution for X?
b. Use your calculator to find the following probabilities:
    i. the probability that 25 adults in the sample prefer saving over spending
    ii. the probability that at most 20 adults prefer saving
    iii. the probability that more than 30 adults prefer saving
c. Using the formulas, calculate the (i) mean and (ii) standard deviation of X.

Solution:

a. $X \sim B(50, 0.6)$
b. Using the TI-83, 83+, 84 calculator with instructions as provided in [link]:
    i. $P(x = 25)$ = binompdf(50, 0.6, 25) = 0.0405
    ii. $P(x \leq 20)$ = binomcdf(50, 0.6, 20) = 0.0034
    iii. $P(x > 30)$ = 1 - binomcdf(50, 0.6, 30) = 1 – 0.5535 = 0.4465
c.
    i. Mean = np = 50(0.6) = 30
    ii. Standard Deviation = $\sqrt{npq} = \sqrt{50\,(0.6)\,(0.4)} \approx 3.4641$

Example:

The lifetime risk of developing pancreatic cancer is about one in 78 (1.28%). Suppose we randomly sample 200 people. Let X = the number of people who will develop pancreatic cancer.

Exercise:

Problem:

a. What is the probability distribution for X?
b. Using the formulas, calculate the (i) mean and (ii) standard deviation of X.
c. Use your calculator to find the probability that at most eight people develop pancreatic cancer
d. Is it more likely that five or six people will develop pancreatic cancer? Justify your answer numerically.

Solution:

a. $X \sim B(200, 0.0128)$

b.
   i. Mean = np = 200(0.0128) = 2.56
   ii. Standard Deviation = $\sqrt{npq} = \sqrt{(200)(0.0128)(0.9872)} \approx 1.5897$

c. Using the TI-83, 83+, 84 calculator with instructions as provided in [link]:

   $P(x \leq 8)$ = binomcdf(200, 0.0128, 8) = 0.9988

d. $P(x = 5)$ = binompdf(200, 0.0128, 5) = 0.0707

   $P(x = 6)$ = binompdf(200, 0.0128, 6) = 0.0298

   So $P(x = 5) > P(x = 6)$; it is more likely that five people will develop cancer than six.

Note:
Try It
Exercise:

Problem:

During the 2013 regular NBA season, DeAndre Jordan of the Los Angeles Clippers had the highest field goal completion rate in the league. DeAndre scored with 61.3% of his shots. Suppose you choose a random sample of 80 shots made by DeAndre during the 2013 season. Let X = the number of shots that scored points.

a. What is the probability distribution for X?
b. Using the formulas, calculate the (i) mean and (ii) standard deviation of X.
c. Use your calculator to find the probability that DeAndre scored with 60 of these shots.
d. Find the probability that DeAndre scored with more than 50 of these shots.

Solution:

a. $X \sim B(80, 0.613)$

b.
   i. Mean = np = 80(0.613) = 49.04
   ii. Standard Deviation = $\sqrt{npq} = \sqrt{80(0.613)(0.387)} \approx 4.3564$

c. Using the TI-83, 83+, 84 calculator with instructions as provided in [link]:

   $P(x = 60)$ = binompdf(80, 0.613, 60) = 0.0036
d. $P(x > 50) = 1 - P(x \le 50)$ = 1 – binomcdf(80, 0.613, 50) = 1 – 0.6282 = 0.3718

Example:
The following example illustrates a problem that is **not** binomial. It violates the condition of independence.

ABC College has a student advisory committee made up of ten staff members and six students. The committee wishes to choose a chairperson and a recorder. What is the probability that the chairperson and recorder are both students?

The names of all committee members are put into a box, and two names are drawn **without replacement**. The first name drawn determines the chairperson and the second name the recorder. There are two trials. However, the trials are not independent because the outcome of the first trial affects the outcome of the second trial.

The probability of a student on the first draw is $\frac{6}{16}$. The probability of a student on the second draw is $\frac{5}{15}$, when the first draw selects a student. The probability is $\frac{6}{15}$, when the first draw selects a staff member. The probability of drawing a student's name changes for each of the trials and, therefore, violates the condition of independence.

Note:
Try It
Exercise:

Problem:

A lacrosse team is selecting a captain. The names of all the seniors are put into a hat, and the first three that are drawn will be the captains. The names are not replaced once they are drawn (one person cannot be two captains). You want to see if the captains all play the same position. State whether this is binomial or not and state why.

Solution:

This is not binomial because the names are not replaced, which means the probability changes for each time a name is drawn. This violates the condition of independence.

References

“Access to electricity (% of population),” The World Bank, 2013. Available online at http://data.worldbank.org/indicator/EG.ELC.ACCS.ZS?order=wbapi_data_value_2009%20wbapi_data_value%20wbapi_data_value-first&sort=asc (accessed May 15, 2015).

“Distance Education.” Wikipedia. Available online at http://en.wikipedia.org/wiki/Distance_education (accessed May 15, 2013).

“NBA Statistics – 2013,” ESPN NBA, 2013. Available online at http://espn.go.com/nba/statistics/_/seasontype/2 (accessed May 15, 2013).

Newport, Frank. “Americans Still Enjoy Saving Rather than Spending: Few demographic differences seen in these views other than by income,” GALLUP® Economy, 2013. Available online at http://www.gallup.com/poll/162368/americans-enjoy-saving-rather-spending.aspx (accessed May 15, 2013).

Pryor, John H., Linda DeAngelo, Laura Palucki Blake, Sylvia Hurtado, Serge Tran. *The American Freshman: National Norms Fall 2011*. Los Angeles: Cooperative Institutional Research Program at the Higher Education Research Institute at UCLA, 2011. Also available online at http://heri.ucla.edu/PDFs/pubs/TFS/Norms/Monographs/TheAmericanFreshman2011.pdf (accessed May 15, 2013).

“The World FactBook,” Central Intelligence Agency. Available online at https://www.cia.gov/library/publications/the-world-factbook/geos/af.html (accessed May 15, 2013).

“What are the key statistics about pancreatic cancer?” American Cancer Society, 2013. Available online at

http://www.cancer.org/cancer/pancreaticcancer/detailedguide/pancreatic-cancer-key-statistics (accessed May 15, 2013).

Section Review

A statistical experiment can be classified as a binomial experiment if the following conditions are met:

1. There are a fixed number of trials, n.
2. There are only two possible outcomes, called "success" and, "failure" for each trial. The letter p denotes the probability of a success on one trial and q denotes the probability of a failure on one trial.
3. The n trials are independent and are repeated using identical conditions.

The outcomes of a binomial experiment fit a binomial probability distribution. The random variable X = the number of successes obtained in the n independent trials. The mean of X can be calculated using the formula μ = np, and the standard deviation is given by the formula σ = $\sqrt{npq}$.

Formula Review

X ~ $B(n, p)$ means that the discrete random variable X has a binomial probability distribution with n trials and probability of success p.

X = the number of successes in n independent trials

n = the number of independent trials

X takes on the values x = 0, 1, 2, 3, ..., n

p = the probability of a success for any trial

q = the probability of a failure for any trial

p + $q = 1$

$q = 1 - p$

The mean of X is $\mu = np$. The standard deviation of X is $\sigma = \sqrt{npq}$.

Use the following information to answer the next eight exercises: The Higher Education Research Institute at UCLA collected data from 203,967 incoming first-time, full-time freshmen from 270 four-year colleges and universities in the U.S. 71.3% of those students replied that, yes, they believe that same-sex couples should have the right to legal marital status. Suppose that you randomly pick eight first-time, full-time freshmen from the survey. You are interested in the number that believes that same sex-couples should have the right to legal marital status.

Exercise:

Problem: In words, define the random variable X.

Solution:

X = the number that reply "yes"

Exercise:

Problem: X ~ _____(_____,_____)

Exercise:

Problem: What values does the random variable X take on?

Solution:

0, 1, 2, 3, 4, 5, 6, 7, 8

Exercise:

Problem: Construct the probability distribution function (PDF).

x	$P(x)$

Exercise:

Problem:

On average (μ), how many would you expect to answer yes?

Solution:

5.7

Exercise:

Problem: What is the standard deviation (σ)?

Exercise:

Problem:

What is the probability that at most five of the freshmen reply “yes”?

Solution:

0.4151

Exercise:

Problem:

What is the probability that at least two of the freshmen reply “yes”?

HOMEWORK

Exercise:

Problem:

According to a recent article the average number of babies born with significant hearing loss (deafness) is approximately two per 1,000 babies in a healthy baby nursery. The number climbs to an average of 30 per 1,000 babies in an intensive care nursery.

Suppose that 1,000 babies from healthy baby nurseries were randomly surveyed. Find the probability that exactly two babies were born deaf.

Use the following information to answer the next four exercises. Recently, a nurse commented that when a patient calls the medical advice line claiming to have the flu, the chance that he or she truly has the flu (and not just a nasty cold) is only about 4%. Of the next 25 patients calling in claiming to have the flu, we are interested in how many actually have the flu.

Exercise:

Problem: Define the random variable and list its possible values.

Solution:

X = the number of patients calling in claiming to have the flu, who actually have the flu.

x = 0, 1, 2, ...25

Exercise:

Problem: State the distribution of X.

Exercise:

Problem:

Find the probability that at least four of the 25 patients actually have the flu.

Solution:

0.0165

Exercise:

Problem:

On average, for every 25 patients calling in, how many do you expect to have the flu?

Exercise:

Problem:

People visiting video rental stores often rent more than one DVD at a time. The probability distribution for DVD rentals per customer at Video To Go is given [link]. There is five-video limit per customer at this store, so nobody ever rents more than five DVDs.

x	$P(x)$
0	0.03
1	0.50
2	0.24

x	$P(x)$
3	
4	0.07
5	0.04

a. Describe the random variable X in words.
b. Find the probability that a customer rents three DVDs.
c. Find the probability that a customer rents at least four DVDs.
d. Find the probability that a customer rents at most two DVDs.

Solution:

a. X = the number of DVDs a Video to Go customer rents
b. 0.12
c. 0.11
d. 0.77

Exercise:

Problem:

A school newspaper reporter decides to randomly survey 12 students to see if they will attend Tet (Vietnamese New Year) festivities this year. Based on past years, she knows that 18% of students attend Tet festivities. We are interested in the number of students who will attend the festivities.

a. In words, define the random variable X.
b. List the values that X may take on.
c. Give the distribution of X. X ~ _____(_____,_____)
d. How many of the 12 students do we expect to attend the festivities?
e. Find the probability that at most four students will attend.

f. Find the probability that more than two students will attend.

Use the following information to answer the next two exercises: The probability that the San Jose Sharks will win any given game is 0.3694 based on a 13-year win history of 382 wins out of 1,034 games played (as of a certain date). An upcoming monthly schedule contains 12 games.
Exercise:

Problem: The expected number of wins for that upcoming month is:

a. 1.67
b. 12
c. $\frac{382}{1043}$
d. 4.43

Solution:

d. 4.43

Let X = the number of games won in that upcoming month.
Exercise:

Problem:

What is the probability that the San Jose Sharks win six games in that upcoming month?

a. 0.1476
b. 0.2336
c. 0.7664
d. 0.8903

Exercise:

Problem:

What is the probability that the San Jose Sharks win at least five games in that upcoming month

a. 0.3694
b. 0.5266
c. 0.4734
d. 0.2305

Solution:

c

Exercise:

Problem:

A student takes a ten-question true-false quiz, but did not study and randomly guesses each answer. Find the probability that the student passes the quiz with a grade of at least 70% of the questions correct.

Exercise:

Problem:

A student takes a 32-question multiple-choice exam, but did not study and randomly guesses each answer. Each question has three possible choices for the answer. Find the probability that the student guesses **more than** 75% of the questions correctly.

Solution:

- X = number of questions answered correctly
- $X \sim B\left(32, \frac{1}{3}\right)$
- We are interested in MORE THAN 75% of 32 questions correct. 75% of 32 is 24. We want to find $P(x > 24)$. The event "more than 24" is the complement of "less than or equal to 24."

- Using your calculator's distribution menu: 1 – binomcdf $\left(32, \frac{1}{3}, 24\right)$
- $P(x > 24) = 0$
- The probability of getting more than 75% of the 32 questions correct when randomly guessing is very small and practically zero.

Exercise:

Problem:

Six different colored dice are rolled. Of interest is the number of dice that show a one.

a. In words, define the random variable X.
b. List the values that X may take on.
c. Give the distribution of X. X ~ _____(_____,_____)
d. On average, how many dice would you expect to show a one?
e. Find the probability that all six dice show a one.
f. Is it more likely that three or that four dice will show a one? Use numbers to justify your answer numerically.

Exercise:

Problem:

More than 96 percent of the very largest colleges and universities (more than 15,000 total enrollments) have some online offerings. Suppose you randomly pick 13 such institutions. We are interested in the number that offer distance learning courses.

a. In words, define the random variable X.
b. List the values that X may take on.
c. Give the distribution of X. X ~ _____(_____,_____)
d. On average, how many schools would you expect to offer such courses?
e. Find the probability that at most ten offer such courses.

f. Is it more likely that 12 or that 13 will offer such courses? Use numbers to justify your answer numerically and answer in a complete sentence.

Solution:

a. X = the number of college and universities that offer online offerings.
b. 0, 1, 2, …, 13
c. $X \sim B(13, 0.96)$
d. 12.48
e. 0.0135
f. $P(x = 12) = 0.3186$ $P(x = 13) = 0.5882$ More likely to get 13.

Exercise:

Problem:

Suppose that about 85% of graduating students attend their graduation. A group of 22 graduating students is randomly chosen.

a. In words, define the random variable X.
b. List the values that X may take on.
c. Give the distribution of X. X ~ _____(_____,_____)
d. How many are expected to attend their graduation?
e. Find the probability that 17 or 18 attend.
f. Based on numerical values, would you be surprised if all 22 attended graduation? Justify your answer numerically.

Exercise:

Problem:

At The Fencing Center, 60% of the fencers use the foil as their main weapon. We randomly survey 25 fencers at The Fencing Center. We are interested in the number of fencers who do **not** use the foil as their main weapon.

a. In words, define the random variable X.
b. List the values that X may take on.
c. Give the distribution of X. X ~ _____(_____,_____)
d. How many are expected to **not** to use the foil as their main weapon?
e. Find the probability that six do **not** use the foil as their main weapon.
f. Based on numerical values, would you be surprised if all 25 did **not** use foil as their main weapon? Justify your answer numerically.

Solution:

a. X = the number of fencers who do **not** use the foil as their main weapon
b. 0, 1, 2, 3,... 25
c. $X \sim B(25{,}0.40)$
d. 10
e. 0.0442
f. The probability that all 25 not use the foil is almost zero. Therefore, it would be very surprising.

Exercise:

Problem:

Approximately 8% of students at a local high school participate in after-school sports all four years of high school. A group of 60 seniors is randomly chosen. Of interest is the number who participated in after-school sports all four years of high school.

a. In words, define the random variable X.
b. List the values that X may take on.
c. Give the distribution of X. X ~ _____(_____,_____)
d. How many seniors are expected to have participated in after-school sports all four years of high school?

e. Based on numerical values, would you be surprised if none of the seniors participated in after-school sports all four years of high school? Justify your answer numerically.
f. Based upon numerical values, is it more likely that four or that five of the seniors participated in after-school sports all four years of high school? Justify your answer numerically.

Exercise:

Problem:

The chance of an IRS audit for a tax return with over $25,000 in income is about 2% per year. We are interested in the expected number of audits a person with that income has in a 20-year period. Assume each year is independent.

a. In words, define the random variable X.
b. List the values that X may take on.
c. Give the distribution of X. X ~ _____(_____,_____)
d. How many audits are expected in a 20-year period?
e. Find the probability that a person is not audited at all.
f. Find the probability that a person is audited more than twice.

Solution:

a. X = the number of audits in a 20-year period
b. 0, 1, 2, …, 20
c. X ~ $B(20, 0.02)$
d. 0.4
e. 0.6676
f. 0.0071

Exercise:

Problem:

It has been estimated that only about 30% of California residents have adequate earthquake supplies. Suppose you randomly survey 11 California residents. We are interested in the number who have adequate earthquake supplies.

a. In words, define the random variable X.
b. List the values that X may take on.
c. Give the distribution of X. X ~ _____(_____,_____)
d. What is the probability that at least eight have adequate earthquake supplies?
e. Is it more likely that none or that all of the residents surveyed will have adequate earthquake supplies? Why?
f. How many residents do you expect will have adequate earthquake supplies?

Exercise:

Problem:

There are two similar games played for Chinese New Year and Vietnamese New Year. In the Chinese version, fair dice with numbers 1, 2, 3, 4, 5, and 6 are used, along with a board with those numbers. In the Vietnamese version, fair dice with pictures of a gourd, fish, rooster, crab, crayfish, and deer are used. The board has those six objects on it, also. We will play with bets being $1. The player places a bet on a number or object. The “house” rolls three dice. If none of the dice show the number or object that was bet, the house keeps the $1 bet. If one of the dice shows the number or object bet (and the other two do not show it), the player gets back his or her $1 bet, plus $1 profit. If two of the dice show the number or object bet (and the third die does not show it), the player gets back his or her $1 bet, plus $2 profit. If all three dice show the number or object bet, the player gets back his or her $1 bet, plus $3 profit. Let X = number of matches and Y = profit per game.

a. In words, define the random variable X.
b. List the values that X may take on.
c. Give the distribution of X. X ~ _____(_____,_____)
d. List the values that Y may take on. Then, construct one PDF table that includes both X and Y and their probabilities.
e. Calculate the average expected matches over the long run of playing this game for the player.
f. Calculate the average expected earnings over the long run of playing this game for the player.
g. Determine who has the advantage, the player or the house.

Solution:

1. X = the number of matches
2. 0, 1, 2, 3
3. $X \sim B\left(3, \frac{1}{6}\right)$
4. In dollars: −1, 1, 2, 3
5. $\frac{1}{2}$
6. Multiply each Y value by the corresponding X probability from the PDF table. The answer is −0.0787. You lose about eight cents, on average, per game.
7. The house has the advantage.

Exercise:

Problem:

According to The World Bank, only 9% of the population of Uganda had access to electricity as of 2009. Suppose we randomly sample 150 people in Uganda. Let X = the number of people who have access to electricity.

a. What is the probability distribution for X?
b. Using the formulas, calculate the mean and standard deviation of X.

c. Use your calculator to find the probability that 15 people in the sample have access to electricity.
d. Find the probability that at most ten people in the sample have access to electricity.
e. Find the probability that more than 25 people in the sample have access to electricity.

Exercise:

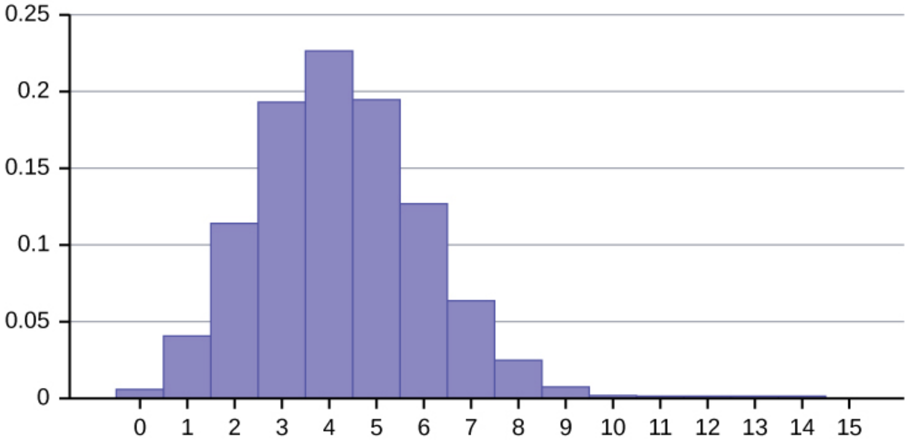
Problem:

The literacy rate for a nation measures the proportion of people age 15 and over that can read and write. The literacy rate in Afghanistan is 28.1%. Suppose you choose 15 people in Afghanistan at random. Let X = the number of people who are literate.

a. Sketch a graph of the probability distribution of X.
b. Using the formulas, calculate the (i) mean and (ii) standard deviation of X.
c. Find the probability that more than five people in the sample are literate. Is it is more likely that three people or four people are literate.

Solution:

a. $X \sim B(15, 0.281)$



b. i. Mean = μ = np = 15(0.281) = 4.215
ii. Standard Deviation = σ = $\sqrt{npq}$ = $\sqrt{15(0.281)(0.719)}$ = 1.7409

c. $P(x > 5)$ = 1 – $P(x \le 5)$ = 1 – binomcdf(15, 0.281, 5) = 1 – 0.7754 = 0.2246
$P(x = 3)$ = binompdf(15, 0.281, 3) = 0.1927
$P(x = 4)$ = binompdf(15, 0.281, 4) = 0.2259
It is more likely that four people are literate that three people are.

Glossary

Binomial Experiment

a statistical experiment that satisfies the following three conditions:

1. There are a fixed number of trials, n.
2. There are only two possible outcomes, called "success" and, "failure," for each trial. The letter p denotes the probability of a success on one trial, and q denotes the probability of a failure on one trial.
3. The n trials are independent and are repeated using identical conditions.

Bernoulli Trials

an experiment with the following characteristics:

1. There are only two possible outcomes called “success” and “failure” for each trial.
2. The probability p of a success is the same for any trial (so the probability $q = 1 - p$ of a failure is the same for any trial).

Binomial Probability Distribution

a discrete random variable (RV) that arises from Bernoulli trials; there are a fixed number, n, of independent trials. “Independent” means that the result of any trial (for example, trial one) does not affect the results of the following trials, and all trials are conducted under the same conditions. Under these circumstances the binomial RV X is defined as the number of successes in n trials. The notation is: X ~ $B(n, p)$. The mean is μ = np and the standard deviation is σ = $\sqrt{npq}$.

The Poisson Distribution

There are two main characteristics of a Poisson experiment.

1. The **Poisson probability distribution** gives the probability of a number of events occurring in a **fixed interval** of time or space if these events happen with a **known average rate** and independently of the time since the last event. For example, a book editor might be interested in the number of words spelled incorrectly in a particular book. It might be that, on the average, there are five words spelled incorrectly in 100 pages. The interval is the 100 pages.
2. The Poisson distribution may be used to approximate the binomial if the probability of success is "small" (such as 0.01) and the number of trials is "large" (such as 1,000). You will verify this relationship in the homework exercises. n is the number of trials, and p is the probability of a "success."

The random variable X = the number of occurrences in the interval of interest.

Example:
The average number of loaves of bread put on a shelf in a bakery in a half-hour period is 12. Of interest is the number of loaves of bread put on the shelf in five minutes. The time interval of interest is five minutes.
Let X = the number of loaves of bread put on the shelf in five minutes. If the average number of loaves put on the shelf in 30 minutes (half-hour) is 12, **then the average number of loaves put on the shelf in five minutes is** $\left(\frac{5}{30}\right)(12) = 2$ loaves of bread.

Note:
Try It
Exercise:

Problem:

The average number of fish caught in an hour is eight. Of interest is the number of fish caught in 15 minutes. The time interval of interest is 15 minutes. What is the average number of fish caught in 15 minutes?

Solution:

$\left(\frac{15}{60}\right)(8) = 2$ fish

Example:
Exercise:

Problem:

A bank expects to receive six bad checks per day, on average. What is the probability of the bank getting fewer than five bad checks on any given day? Of interest is the number of checks the bank receives in one day, so the time interval of interest is one day. Let X = the number of bad checks the bank receives in one day. If the bank expects to receive six bad checks per day then the average is six checks per day. Write a mathematical statement for the probability question.

Solution:

$P(x < 5)$

Note:
Try It
Exercise:

Problem:

An electronics store expects to have ten returns per day on average. The manager wants to know the probability of the store getting fewer than eight returns on any given day. State the probability question mathematically.

Solution:

$P(x < 8)$

Example:
You notice that a news reporter says "uh," on average, two times per broadcast. What is the probability that the news reporter says "uh" more than two times per broadcast.

This is a Poisson problem because you are interested in knowing the number of times the news reporter says "uh" during a broadcast.

Exercise:

Problem: a. What is the interval of interest?

Solution:

a. one broadcast

Exercise:

Problem:

b. What is the average number of times the news reporter says "uh" during one broadcast?

Solution:

b. 2

Exercise:

Problem: c. Let X = ____________. What values does X take on?

Solution:

c. Let X = the number of times the news reporter says "uh" during one broadcast.
x = 0, 1, 2, 3, ...

Exercise:

Problem: d. The probability question is P(______).

Solution:

d. $P(x > 2)$

Note:
Try It
Exercise:

Problem:

An emergency room at a particular hospital gets an average of five patients per hour. A doctor wants to know the probability that the ER gets more than five patients per hour. Give the reason why this would be a Poisson distribution.

Solution:

This problem wants to find the probability of events occurring in a fixed interval of time with a known average rate. The events are independent.

Notation for the Poisson: P = Poisson Probability Distribution Function

$X \sim P(\mu)$

Read this as "X is a random variable with a Poisson distribution." The parameter is μ (or λ); μ (or λ) = the mean for the interval of interest.

The variance is $\sigma^2 = \mu$, and the standard deviation is $\sigma = \sqrt{\mu}$.

Calculating Probabilities Using the Poisson Distribution

To calculate probabilities using the Poisson Distribution, we will use the poissonpdf and poissoncdf functions on the TI 83+ or 84 calculator. These are found under the 2nd DISTR menu.

To calculate $P(x = \text{number})$: Enter poissonpdf(μ, number).

To calculate $P(x \leq \text{number})$: Enter poissoncdf(μ, number).

To calculate $P(x > \text{number})$: Enter $1-$ poissoncdf(μ, number), since $P(x > \text{number})$ is the complement of $P(x \leq \text{number})$.

Note:
The TI calculators may use λ (the Greek letter lambda) to refer to the mean, rather than μ.

Example:
Leah's answering machine receives about six telephone calls between 8 a.m. and 10 a.m. What is the probability that Leah receives more than one call **in the next 15 minutes?**

Let X = the number of calls Leah receives in 15 minutes. (The **interval of interest** is 15 minutes or $\frac{1}{4}$ hour.)

x = 0, 1, 2, 3, ...

We need to determine μ. If Leah receives, on the average, six telephone calls in two hours, and there are eight 15 minute intervals in two hours, then Leah receives

$\left(\frac{1}{8}\right)(6)$ = 0.75 calls in 15 minutes, on average. So, μ = 0.75 for this problem.

$X \sim P(0.75)$

Find $P(x > 1)$, using the calculator.

Note:
Calculator instructions:

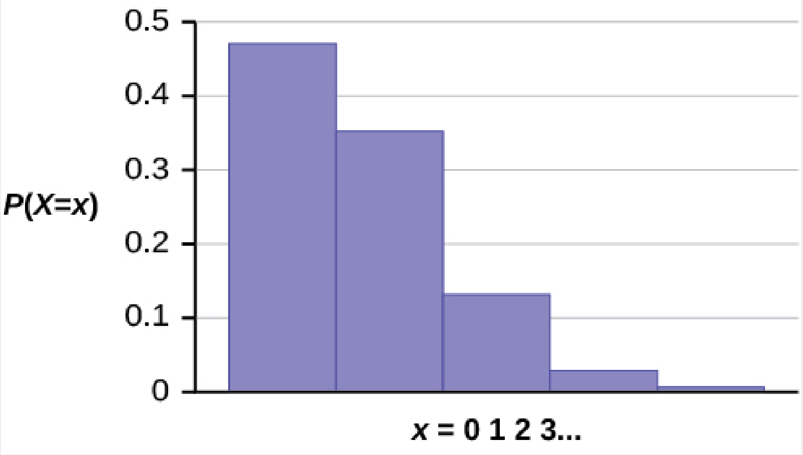
- Press 1 – and then press 2nd DISTR.
- Arrow down to poissoncdf. Press ENTER.
- Enter (.75,1).
- The result is $P(x > 1)$ = 0.1734.

The probability that Leah receives more than one telephone call in the next 15 minutes is about 0.1734:

$P(x > 1) = 1 - \text{poissoncdf}(0.75, 1) = 0.1734$.

The graph of $X \sim P(0.75)$ is:

A Poisson Probability Distribution



Note that, while not all visible, the values of x go on forever.

The y-axis contains the probability of x where X = the number of calls in 15 minutes.

Practice finding probabilities on your calculator by verifying the probabilities shown in the graph above. Note that the first bar represents $x = 0$, the second bar represents $x = 1$, and so on.

Note:

Try It

Exercise:

Problem:

A customer service center receives about ten emails every half-hour. What is the probability that the customer service center receives more than four emails in the next six minutes? Use the TI-83+ or TI-84 calculator to find the answer.

Solution:

$P(x > 4)$ = 0.0527

Example:

According to Baydin, an email management company, an email user gets, on average, 147 emails per day. Let X = the number of emails an email user receives per day. The discrete random variable X takes on the values x = 0, 1, 2 …. The random variable X has a Poisson distribution: $X \sim P(147)$. The mean is 147 emails.

Exercise:

Problem:

a. What is the probability that an email user receives exactly 160 emails per day?
b. What is the probability that an email user receives at most 160 emails per day?
c. What is the standard deviation?

Solution:

a. $P(x = 160)$ = poissonpdf(147, 160) ≈ 0.0180
b. $P(x \leq 160)$ = poissoncdf(147, 160) ≈ 0.8666
c. Standard Deviation = $\sigma = \sqrt{\mu} = \sqrt{147} \approx 12.1244$

Note:
Try It
Exercise:

Problem:

According to a recent poll by the Pew Internet Project, girls between the ages of 14 and 17 send an average of 187 text messages each day. Let X = the number of texts that a girl aged 14 to 17 sends per day. The discrete random variable X takes on the values x = 0, 1, 2 …. The random variable X has a Poisson distribution: X ~ $P(187)$. The mean is 187 text messages.

a. What is the probability that a teen girl sends exactly 175 texts per day?
b. What is the probability that a teen girl sends at most 150 texts per day?
c. What is the standard deviation?

Solution:

a. $P(x = 175)$ = poissonpdf(187, 175) ≈ 0.0203
b. $P(x \le 150)$ = poissoncdf(187, 150) ≈ 0.0030
c. Standard Deviation = $\sigma = \sqrt{\mu} = \sqrt{187} \approx 13.6748$

Example:
Text message users receive or send an average of 41.5 text messages per day.
Exercise:

Problem:

a. How many text messages does a text message user receive or send per hour?

b. What is the probability that a text message user receives or sends two messages per hour?
c. What is the probability that a text message user receives or sends more than two messages per hour?

Solution:

a. Let X = the number of texts that a user sends or receives in one hour. The average number of texts received per hour is $\frac{41.5}{24} \approx$ 1.7292.
b. $X \sim P(1.7292)$, so $P(x = 2)$ = poissonpdf(1.7292, 2) ≈ 0.2653
c. $P(x > 2) = 1 - P(x \leq 2)$ = 1 – poissoncdf(1.7292, 2) ≈ 1 – 0.7495 = 0.2505

Note:
Try It
Exercise:

Problem:

Atlanta’s Hartsfield-Jackson International Airport is the busiest airport in the world. On average there are 2,500 arrivals and departures each day.

a. How many airplanes arrive and depart the airport per hour?
b. What is the probability that there are exactly 100 arrivals and departures in one hour?
c. What is the probability that there are at most 100 arrivals and departures in one hour?

Solution:

a. Let X = the number of airplanes arriving and departing from Hartsfield-Jackson in one hour. The average number of arrivals and departures per hour is $\frac{2{,}500}{24}$ ≈ 104.1667.
b. X ~ $P(104.1667)$, so $P(x = 100)$ = poissonpdf(104.1667, 100) ≈ 0.0366.
c. $P(x \leq 100)$ = poissoncdf(104.1667, 100) ≈ 0.3651.

The Poisson distribution can be used to approximate probabilities for a binomial distribution. This next example demonstrates the relationship between the Poisson and the binomial distributions. Let n represent the number of binomial trials and let p represent the probability of a success for each trial. If n is large enough and p is small enough then the Poisson approximates the binomial very well. In general, n is considered “large enough” if it is greater than or equal to 20. The probability p from the binomial distribution should be less than or equal to 0.05. When the Poisson is used to approximate the binomial, we use the binomial mean $\mu = np$. The variance of X is $\sigma^2 = \mu$ and the standard deviation is $\sigma = \sqrt{\mu}$. The Poisson approximation to a binomial distribution was commonly used in the days before technology made both values very easy to calculate.

Example:
Exercise:

Problem:

On May 13, 2013, starting at 4:30 PM, the probability of low seismic activity for the next 48 hours in Alaska was reported as about 1.02%. Use this information for the next 200 days to find the probability that there will be low seismic activity in ten of the next 200 days. Use both the binomial and Poisson distributions to calculate the probabilities. Are they close?

Solution:

Let X = the number of days with low seismic activity.

Using the binomial distribution:

$P(x = 10)$ = binompdf(200, .0102, 10) ≈ 0.000039

Using the Poisson distribution:

Calculate $\mu = np$ = 200(0.0102) ≈ 2.04
$P(x = 10)$ = poissonpdf(2.04, 10) ≈ 0.000045

We expect the approximation to be good because n is large (greater than 20) and p is small (less than 0.05). The results are close — both probabilities reported are almost 0.

Note:
Try It
Exercise:

Problem:

On May 13, 2013, starting at 4:30 PM, the probability of moderate seismic activity for the next 48 hours in the Kuril Islands off the coast of Japan was reported at about 1.43%. Use this information for the next 100 days to find the probability that there will be low seismic activity in five of the next 100 days. Use both the binomial and Poisson distributions to calculate the probabilities. Are they close?

Solution:

Let X = the number of days with moderate seismic activity.

Using the binomial distribution: $P(x = 5)$ = binompdf(100, 0.0143, 5) ≈ 0.0115

Using the Poisson distribution:

Calculate $\mu = np = 100(0.0143) = 1.43$
$P(x = 5) =$ poissonpdf(1.43, 5) = 0.0119

We expect the approximation to be good because n is large (greater than 20) and p is small (less than 0.05). The results are close — the difference between the values is 0.0004.

Note: While, in some cases, the Poisson can be used to approximate the binomial distribution, it's important to understand the differences between the two distributions. Namely,

- In a binomial experiment, there are a fixed number of trials. However, in a Poisson experiment, there could be an unlimited number of trials. For example, if you are interested in how many times a six is rolled when a die is rolled 36 times, then this is a binomial experiment, since 36 is a fixed number of trials. In contrast, if you are interested in how many times a six is rolled when a person rolls a die continuously for 2 hours, then this would be a Poisson experiment, since, theoretically, there is no limit to how many times a person can roll a die in 2 hours.
- In a binomial experiment, there are a fixed number of outcomes possible, but in a Poisson experiment the number of possible outcomes is limitless. For example, if a die is rolled 36 times (binomial), the number of sixes that are rolled can range from 0 to 36. However, if a die is rolled continuously for 2 hours (Poisson), one cannot say with certainty the maximum number of sixes that can be rolled.
- The binomial distribution is determined by two parameters: p, the probability of success, and n, the number of trials. However, the Poisson distribution is determined by only one parameter: μ, the average rate at which an event occurs.

References

"ATL Fact Sheet," Department of Aviation at the Hartsfield-Jackson Atlanta International Airport, 2013. Available online at http://www.atlanta-airport.com/Airport/ATL/ATL_FactSheet.aspx (accessed May 15, 2013).

Center for Disease Control and Prevention. "Teen Drivers: Fact Sheet," Injury Prevention & Control: Motor Vehicle Safety, October 2, 2012. Available online at http://www.cdc.gov/Motorvehiclesafety/Teen_Drivers/teendrivers_factsheet.html (accessed May 15, 2013).

"Children and Childrearing," Ministry of Health, Labour, and Welfare. Available online at http://www.mhlw.go.jp/english/policy/children/children-childrearing/index.html (accessed May 15, 2013).

"Eating Disorder Statistics," South Carolina Department of Mental Health, 2006. Available online at http://www.state.sc.us/dmh/anorexia/statistics.htm (accessed May 15, 2013).

"Giving Birth in Manila: The maternity ward at the Dr Jose Fabella Memorial Hospital in Manila, the busiest in the Philippines, where there is an average of 60 births a day," theguardian, 2013. Available online at http://www.theguardian.com/world/gallery/2011/jun/08/philippines-health#/?picture=375471900&index=2 (accessed May 15, 2013).

"How Americans Use Text Messaging," Pew Internet, 2013. Available online at http://pewinternet.org/Reports/2011/Cell-Phone-Texting-2011/Main-Report.aspx (accessed May 15, 2013).

Lenhart, Amanda. "Teens, Smartphones & Testing: Texting volum is up while the frequency of voice calling is down. About one in four teens say they own smartphones," Pew Internet, 2012. Available online at http://www.pewinternet.org/~/media/Files/Reports/2012/PIP_Teens_Smartphones_and_Texting.pdf (accessed May 15, 2013).

"One born every minute: the maternity unit where mothers are THREE to a bed," MailOnline. Available online at http://www.dailymail.co.uk/news/article-2001422/Busiest-maternity-ward-planet-averages-60-babies-day-mothers-bed.html (accessed May 15, 2013).

Vanderkam, Laura. “Stop Checking Your Email, Now.” CNNMoney, 2013. Available online at http://management.fortune.cnn.com/2012/10/08/stop-checking-your-email-now/ (accessed May 15, 2013).

“World Earthquakes: Live Earthquake News and Highlights,” World Earthquakes, 2012. http://www.world-earthquakes.com/index.php?option=ethq_prediction (accessed May 15, 2013).

Section Review

A **Poisson probability distribution** of a discrete random variable gives the probability of a number of events occurring in a fixed interval of time or space, if these events happen at a known average rate and independently of the time since the last event. The Poisson distribution may be used to approximate the binomial, if the probability of success is "small" (less than or equal to 0.05) and the number of trials is "large" (greater than or equal to 20).

Formula Review

X ~ $P(\mu)$ means that X has a Poisson probability distribution where X = the number of occurrences in the interval of interest.

X takes on the values x = 0, 1, 2, 3, ...

The mean μ is typically given.

The variance is σ^2 = μ, and the standard deviation is $\sigma = \sqrt{\mu}$.

When $P(\mu)$ is used to approximate a binomial distribution, μ = np where n represents the number of independent trials and p represents the probability of success in a single trial.

Use the following information to answer the next six exercises: On average, a clothing store gets 120 customers per day.

Exercise:

Problem:

Assume the event occurs independently in any given day. Define the random variable X.

Exercise:

Problem: What values does X take on?

Solution:

0, 1, 2, 3, 4, …

Exercise:

Problem: What is the probability of getting 150 customers in one day?

Exercise:

Problem:

What is the probability of getting 35 customers in the first four hours? Assume the store is open 12 hours each day.

Solution:

0.0485

Exercise:

Problem:

What is the probability that the store will have more than 12 customers in the first hour?

Exercise:

Problem:

What is the probability that the store will have fewer than 12 customers in the first two hours?

Solution:

0.0214

Exercise:

Problem:

Which type of distribution can the Poisson model be used to approximate? When would you do this?

Use the following information to answer the next six exercises: On average, eight teens in the U.S. die from motor vehicle injuries per day. As a result, states across the country are debating raising the driving age.

Exercise:

Problem:

Assume the event occurs independently in any given day. In words, define the random variable X.

Solution:

X = the number of U.S. teens who die from motor vehicle injuries per day.

Exercise:

Problem:X ~ _____(__________)

Exercise:

Problem: What values does X take on?

Solution:

0, 1, 2, 3, 4, ...

Exercise:

Problem:

For the given values of the random variable X, fill in the corresponding probabilities.

Exercise:

Problem:

Is it likely that there will be no teens killed from motor vehicle injuries on any given day in the U.S? Justify your answer numerically.

Solution:

No

Exercise:

Problem:

Is it likely that there will be more than 20 teens killed from motor vehicle injuries on any given day in the U.S.? Justify your answer numerically.

HOMEWORK

Exercise:

Problem:

The switchboard in a Minneapolis law office gets an average of 5.5 incoming phone calls during the noon hour on Mondays. Experience shows that the existing staff can handle up to six calls in an hour. Let X = the number of calls received at noon.

a. Find the mean and standard deviation of X.
b. What is the probability that the office receives at most six calls at noon on Monday?
c. Find the probability that the law office receives six calls at noon. What does this mean to the law office staff who get, on average, 5.5 incoming phone calls at noon?
d. What is the probability that the office receives more than eight calls at noon?

Solution:

a. $X \sim P(5.5)$; $\mu = 5.5$; $\sigma = \sqrt{5.5} \approx 2.3452$
b. $P(x \leq 6) = \text{poissoncdf}(5.5, 6) \approx 0.6860$
c. There is a 15.7% probability that the law staff will receive more calls than they can handle.
d. $P(x > 8) = 1 - P(x \leq 8) = 1 - \text{poissoncdf}(5.5, 8) \approx 1 - 0.8944 = 0.1056$

Exercise:

Problem:

The maternity ward at Dr. Jose Fabella Memorial Hospital in Manila in the Philippines is one of the busiest in the world with an average of 60 births per day. Let X = the number of births in an hour.

a. Find the mean and standard deviation of X.
b. Sketch a graph of the probability distribution of X.
c. What is the probability that the maternity ward will deliver three babies in one hour?
d. What is the probability that the maternity ward will deliver at most three babies in one hour?
e. What is the probability that the maternity ward will deliver more than five babies in one hour?

Exercise:

Problem:

A manufacturer of Christmas tree light bulbs knows that 3% of its bulbs are defective. Find the probability that a string of 100 lights contains at most four defective bulbs using both the binomial and Poisson distributions.

Solution:

Let X = the number of defective bulbs in a string.

Using the Poisson distribution:

- μ = np = 100(0.03) = 3
- X ~ $P(3)$
- $P(x \leq 4)$ = poissoncdf(3, 4) ≈ 0.8153

Using the binomial distribution:

- X ~ $B(100, 0.03)$
- $P(x \leq 4)$ = binomcdf(100, 0.03, 4) ≈ 0.8179

The Poisson approximation is very good—the difference between the probabilities is only 0.0026.

Exercise:

Problem:

The average number of children a Japanese woman has in her lifetime is 1.37. Suppose that one Japanese woman is randomly chosen.

a. In words, define the random variable X.
b. List the values that X may take on.
c. Give the distribution of X. X ~ _____(_____,_____)
d. Find the probability that she has no children.
e. Find the probability that she has fewer children than the Japanese average.

f. Find the probability that she has more children than the Japanese average.

Exercise:

Problem:

The average number of children a Spanish woman has in her lifetime is 1.47. Suppose that one Spanish woman is randomly chosen.

a. In words, define the Random Variable X.
b. List the values that X may take on.
c. Give the distribution of X. X ~ _____(_____,_____)
d. Find the probability that she has no children.
e. Find the probability that she has fewer children than the Spanish average.
f. Find the probability that she has more children than the Spanish average .

Solution:

a. X = the number of children for a Spanish woman
b. 0, 1, 2, 3,...
c. X ~ $P(1.47)$
d. 0.2299
e. 0.5679
f. 0.4321

Exercise:

Problem:

Fertile, female cats produce an average of three litters per year. Suppose that one fertile, female cat is randomly chosen. In one year, find the probability she produces:

a. In words, define the random variable X.

b. List the values that X may take on.
c. Give the distribution of X. X ~ _______
d. Find the probability that she has no litters in one year.
e. Find the probability that she has at least two litters in one year.
f. Find the probability that she has exactly three litters in one year.

Exercise:

Problem:

The chance of having an extra fortune in a fortune cookie is about 3%. Given a bag of 144 fortune cookies, we are interested in the number of cookies with an extra fortune. Two distributions may be used to solve this problem, but only use one distribution to solve the problem.

a. In words, define the random variable X.
b. List the values that X may take on.
c. Give the distribution of X. X ~ _____(_____,_____)
d. How many cookies do we expect to have an extra fortune?
e. Find the probability that none of the cookies have an extra fortune.
f. Find the probability that more than three have an extra fortune.
g. As n increases, what happens involving the probabilities using the two distributions? Explain in complete sentences.

Solution:

a. X = the number of fortune cookies that have an extra fortune
b. 0, 1, 2, 3,... 144
c. $X \sim B(144, 0.03)$ or $P(4.32)$
d. 4.32
e. 0.0124 or 0.0133
f. 0.6300 or 0.6264
g. As n gets larger, the probabilities get closer together.

Exercise:

Problem:

According to the South Carolina Department of Mental Health web site, for every 200 U.S. women, the average number who suffer from anorexia is one. Out of a randomly chosen group of 600 U.S. women determine the following.

a. In words, define the random variable X.
b. List the values that X may take on.
c. Give the distribution of X. X ~ _____(_____,_____)
d. How many are expected to suffer from anorexia?
e. Find the probability that no one suffers from anorexia.
f. Find the probability that more than four suffer from anorexia.

Exercise:

Problem:

The chance of an IRS audit for a tax return with over $25,000 in income is about 2% per year. Suppose that 100 people with tax returns over $25,000 are randomly picked. We are interested in the number of people audited in one year. Use a Poisson distribution to anwer the following questions.

a. In words, define the random variable X.
b. List the values that X may take on.
c. Give the distribution of X. X ~ _____(_____,_____)
d. How many are expected to be audited?
e. Find the probability that no one was audited.
f. Find the probability that at least three were audited.

Solution:

a. X = the number of people audited in one year
b. 0, 1, 2, ..., 100
c. $X \sim P(2)$
d. 2

e. 0.1353
f. 0.3233

Exercise:

Problem:

Approximately 8% of students at a local high school participate in after-school sports all four years of high school. A group of 60 seniors is randomly chosen. Of interest is the number that participated in after-school sports all four years of high school.

a. In words, define the random variable X.
b. List the values that X may take on.
c. Give the distribution of X. X ~ _____(_____,_____)
d. How many seniors are expected to have participated in after-school sports all four years of high school?
e. Based on numerical values, would you be surprised if none of the seniors participated in after-school sports all four years of high school? Justify your answer numerically.
f. Based on numerical values, is it more likely that four or that five of the seniors participated in after-school sports all four years of high school? Justify your answer numerically.

Exercise:

Problem:

On average, Pierre, an amateur chef, drops three pieces of egg shell into every two cake batters he makes. Suppose that you buy one of his cakes.

a. In words, define the random variable X.
b. List the values that X may take on.
c. Give the distribution of X. X ~ _____(_____,_____)
d. On average, how many pieces of egg shell do you expect to be in the cake?

e. What is the probability that there will not be any pieces of egg shell in the cake?
f. Let’s say that you buy one of Pierre’s cakes each week for six weeks. What is the probability that there will not be any egg shell in any of the cakes?
g. Based upon the average given for Pierre, is it possible for there to be seven pieces of shell in the cake? Why?

Solution:

a. X = the number of shell pieces in one cake
b. 0, 1, 2, 3,...
c. $X \sim P(1.5)$
d. 1.5
e. 0.2231
f. 0.0001
g. Yes

Use the following information to answer the next two exercises: The average number of times per week that Mrs. Plum’s cats wake her up at night because they want to play is ten. We are interested in the number of times her cats wake her up each week.

Exercise:

Problem: In words, the random variable X = __________________

a. the number of times Mrs. Plum’s cats wake her up each week.
b. the number of times Mrs. Plum’s cats wake her up each hour.
c. the number of times Mrs. Plum’s cats wake her up each night.
d. the number of times Mrs. Plum’s cats wake her up.

Exercise:

Problem:

Find the probability that her cats will wake her up no more than five times next week.

a. 0.5000
b. 0.9329
c. 0.0378
d. 0.0671

Solution:

d

Glossary

Poisson Probability Distribution

a discrete random variable (RV) that counts the number of times a certain event will occur in a specific interval; characteristics of the variable:

- The probability that the event occurs in a given interval is the same for all intervals.
- The events occur with a known mean and independently of the time since the last event.

The distribution is defined by the mean μ of the event in the interval. Notation: X ~ $P(\mu)$. The mean is μ = np. The standard deviation is $\sigma = \sqrt{\mu}$. The Poisson distribution is often used to approximate the binomial distribution, when n is "large" and p is "small" (a general rule is that n should be greater than or equal to 20 and p should be less than or equal to 0.05).

Continuous Random Variables: Introduction class="introduction"

The heights of these radish plants are continuous random variables. (Credit: Rev Stan)

Note:
Chapter Objectives
By the end of this chapter, the student should be able to:

- Recognize and understand continuous probability density functions in general.
- Recognize the uniform probability distribution and apply it appropriately.
- Recognize the exponential probability distribution and apply it appropriately. (optional)

Continuous random variables have many applications. Baseball batting averages, IQ scores, the length of time a long distance telephone call lasts, the amount of money a person carries, the length of time a computer chip lasts, and SAT scores are just a few. The field of reliability depends on a variety of continuous random variables.

Note:
Note
The values of discrete and continuous random variables can be ambiguous. For example, if X is equal to the number of miles (to the nearest mile) you drive to work, then X is a discrete random variable. You count the miles. If X is the distance you drive to work, then you measure values of X and X is a continuous random variable. For a second example, if X is equal to the number of books in a backpack, then X is a discrete random variable. If X is the weight of a book, then X is a continuous random variable because weights are measured. How the random variable is defined is very important.

Properties of Continuous Probability Distributions

The graph of a continuous probability distribution is a curve. Probability is represented by area under the curve.

The curve is called the **probability density function** (abbreviated as **pdf**). We use the symbol $f(x)$ to represent the curve. $f(x)$ is the function that corresponds to the graph; we use the density function $f(x)$ to draw the graph of the probability distribution.

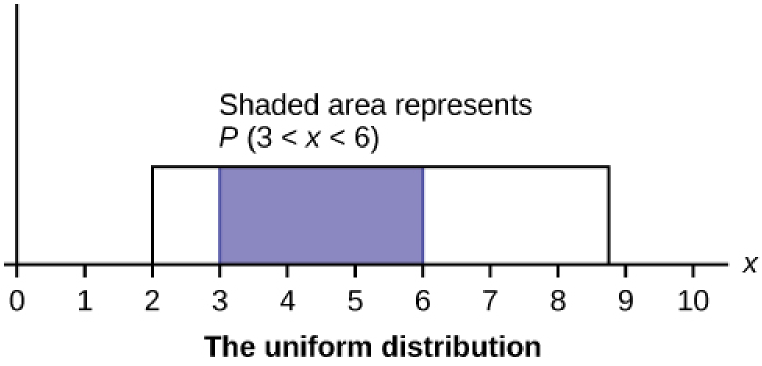
Area under the curve is given by a different function called the **cumulative distribution function** (abbreviated as **cdf**). The cumulative distribution function is used to evaluate probability as area.

- The outcomes are measured, not counted.
- The entire area under the curve and above the x-axis is equal to one.
- Probability is found for intervals of x values rather than for individual x values.
- $P(\mathrm{c} < x < \mathrm{d})$ is the probability that the random variable X is in the interval between the values *c* and *d*. $P(\mathrm{c} < x < \mathrm{d})$ is the area under the curve, above the x-axis, to the right of *c* and the left of *d*.
- $P(x = \mathrm{c}) = 0$ The probability that x takes on any single individual value is zero. The area below the curve, above the x-axis, and between $x = c$ and $x = c$ has no width, and therefore no area (area = 0). Since the probability is equal to the area, the probability is also zero.
- $P(\mathrm{c} < x < \mathrm{d})$ is the same as $P(\mathrm{c} \le x \le \mathrm{d})$ because probability is equal to area.

We will find the area that represents probability by using geometry, formulas, technology, or probability tables. In general, calculus is needed to find the area under the curve for many probability density functions. When we use formulas to find the area in this textbook, the formulas were found by using the techniques of integral calculus. However, because most students taking this course have not studied calculus, we will not be using calculus in this textbook.

There are many continuous probability distributions. When using a continuous probability distribution to model probability, the distribution used is selected to model and fit the particular situation in the best way.

In this chapter and the next, we will study the uniform distribution and the normal distribution. The following graphs illustrate these distributions.



The graph shows a Uniform Distribution with the area between x = 3 and x = 6 shaded to represent the probability that the value of the random variable X is in the interval between three and six.

Glossary

Probability Density Function
: the curve representing the graph of a continuous probability distribution.

Cumulative Distribution Function
: a function which gives the area under a curve and represents a probability.

Uniform Distribution

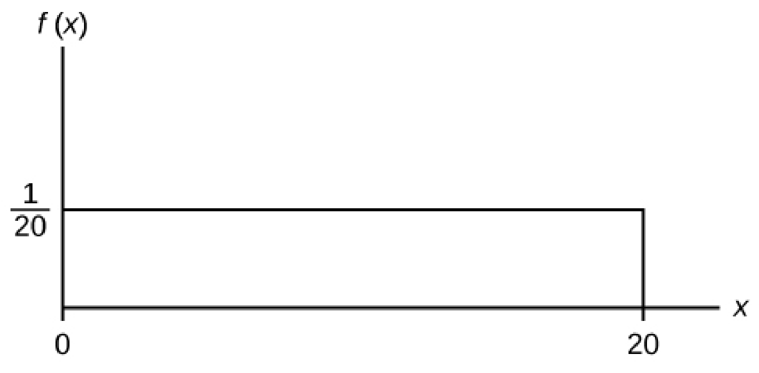
a continuous random variable (RV) that has equally likely outcomes over the domain, $a < x < b$; it is often referred as the **rectangular distribution** because the graph of the pdf has the form of a rectangle. Notation: $X \sim U(a, b)$. The mean is $\mu = \frac{a+b}{2}$ and the standard deviation is $\sigma = \sqrt{\frac{(b-a)^2}{12}}$. The probability density function is $f(x) = \frac{1}{b-a}$ for $a < x < b$ or $a \le x \le b$. The cumulative distribution is $P(X \le x) = \frac{x-a}{b-a}$.

Continuous Probability Functions

We begin by defining a continuous probability density function. We use the function notation $f(x)$. Intermediate algebra may have been your first formal introduction to functions. In the study of probability, the functions we study are special. We define the function $f(x)$ so that the area between it and the x-axis is equal to a probability. Since the maximum probability is one, the maximum area is also one. **For continuous probability distributions, PROBABILITY = AREA.**

Example:

Consider the function $f(x) = \frac{1}{20}$ for $0 \le x \le 20$, where x = a real number. The graph of $f(x) = \frac{1}{20}$ is a horizontal line. However, since $0 \le x \le 20$, $f(x)$ is restricted to the portion between $x = 0$ and $x = 20$, inclusive.



$f(x) = \frac{1}{20}$ **for** $0 \le x \le 20$.

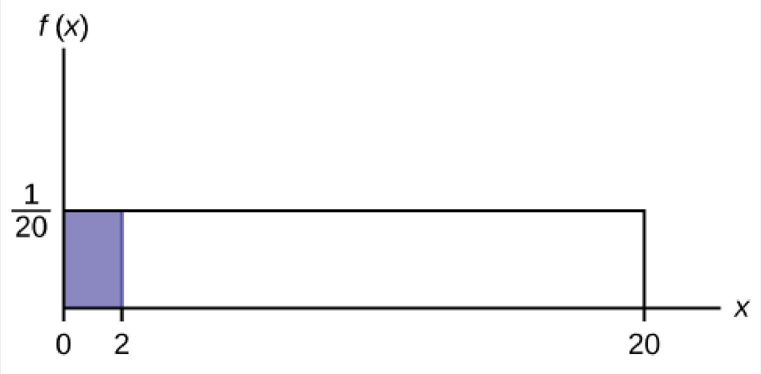
The graph of $f(x) = \frac{1}{20}$ is a horizontal line segment when $0 \le x \le 20$.

The area between $f(x) = \frac{1}{20}$ where $0 \le x \le 20$ and the x-axis is the area of a rectangle with base = 20 and height = $\frac{1}{20}$.

Equation:

$$\text{AREA} = 20\left(\frac{1}{20}\right) = 1$$

Suppose we want to find the area between $f(x) = \frac{1}{20}$ and the x-axis where 0 < x < 2.



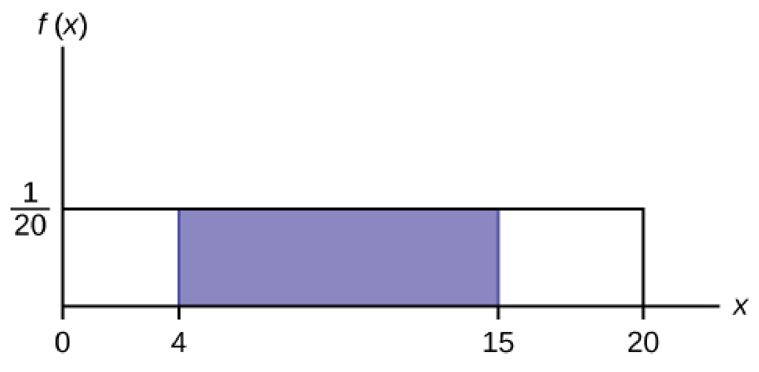
$$\text{AREA} = (2 - 0)\left(\frac{1}{20}\right) = 0.1$$

$$(2 - 0) = 2 = \text{base of a rectangle}$$

Note:
Reminder
area of a rectangle = (base)(height).

The area corresponds to a probability. The probability that x is between zero and two is 0.1, which can be written mathematically as $P(0 < x < 2) = P(x < 2) = 0.1$.

Suppose we want to find the area between $f(x) = \frac{1}{20}$ and the x-axis where 4 < x < 15.



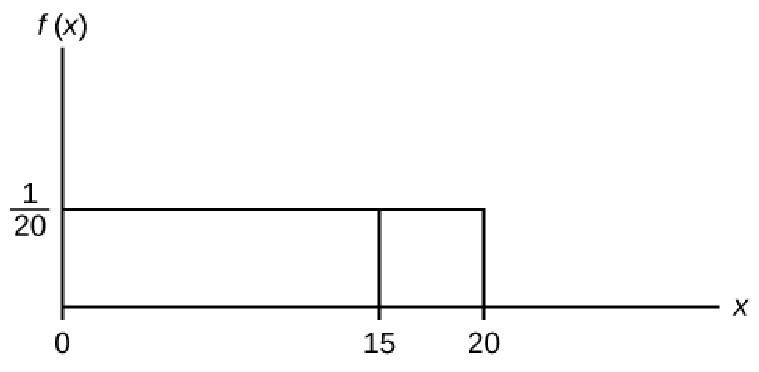
$$\text{AREA} = (15 - 4)\left(\frac{1}{20}\right) = 0.55$$

$$\text{AREA} = (15 - 4)\left(\frac{1}{20}\right) = 0.55$$

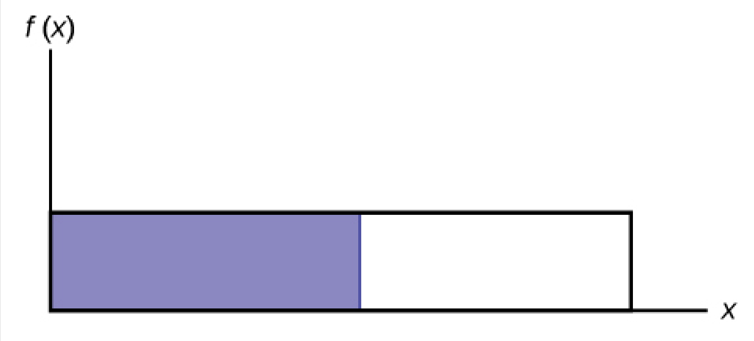
$$(15 - 4) = 11 = \text{the base of a rectangle}$$

The area corresponds to the probability $P(4 < x < 15) = 0.55$.

Suppose we want to find $P(x = 15)$. On an x-y graph, $x = 15$ is a vertical line. A vertical line has no width (or zero width). Therefore, $P(x = 15) =$ (base)(height) = $(0)\left(\frac{1}{20}\right) = 0$

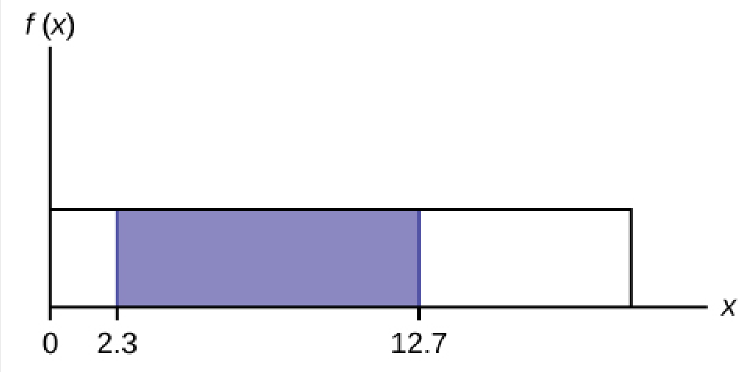


$P(X \leq x)$ (can be written as $P(X < x)$ for continuous distributions) is called the cumulative distribution function or CDF. Notice the "less than or equal to" symbol. We can use the CDF to calculate $P(X > x)$. The CDF gives "area to the left" and $P(X > x)$ gives "area to the right". Find $P(X > x)$ for continuous distributions as follows: $P(X > x) = 1 - P(X < x)$.



Label the graph with $f(x)$ and x. Scale the x and y axes with the maximum x and y values. $f(x) = \frac{1}{20}$, $0 \le x \le 20$.

To calculate the probability that x is between two values, look at the following graph. Shade the region between $x = 2.3$ and $x = 12.7$. Then calculate the shaded area of a rectangle.



$$P(2.3 < x < 12.7) = (\text{base})(\text{height}) = (12.7 - 2.3)\left(\frac{1}{20}\right) = 0.52$$

Note:

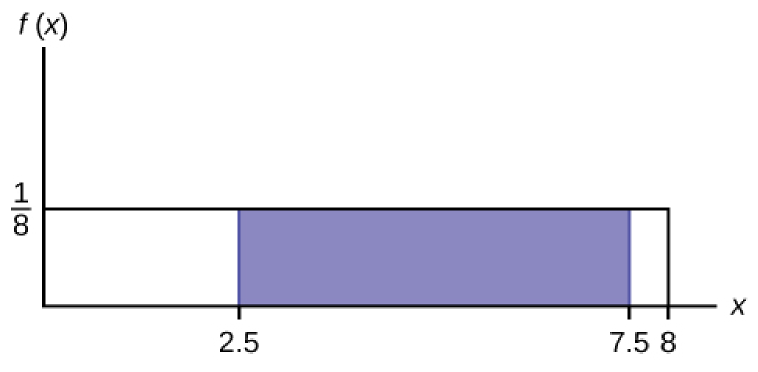
Try It

Exercise:

Problem:

Consider the function $f(x) = \frac{1}{8}$ for $0 \le x \le 8$. Draw the graph of $f(x)$ and find $P(2.5 < x < 7.5)$.

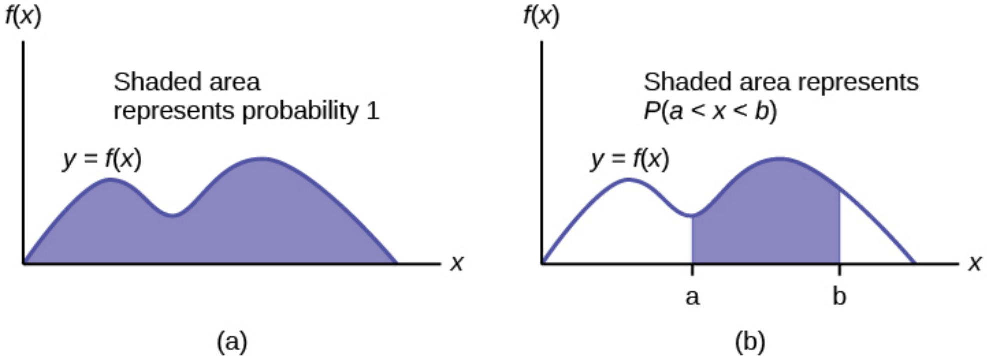
Solution:



$P(2.5 < x < 7.5) = 0.625$

Section Review

The probability density function (pdf) is used to describe probabilities for continuous random variables. The area under the density curve between two points corresponds to the probability that the variable falls between those two values. In other words, the area under the density curve between points a and b is equal to $P(\mathrm{a} < x < \mathrm{b})$. The cumulative distribution function (cdf) gives the probability as an area. If X is a continuous random variable, the probability density function (pdf), $f(x)$, is used to draw the graph of the probability distribution. The total area under the graph of $f(x)$ is one. The area under the graph of $f(x)$ and between values a and b gives the probability $P(\mathrm{a} < x < \mathrm{b})$.



The cumulative distribution function (cdf) of X is defined by $P(X \leq x)$. It is a function of x that gives the probability that the random variable is less than or equal to x.

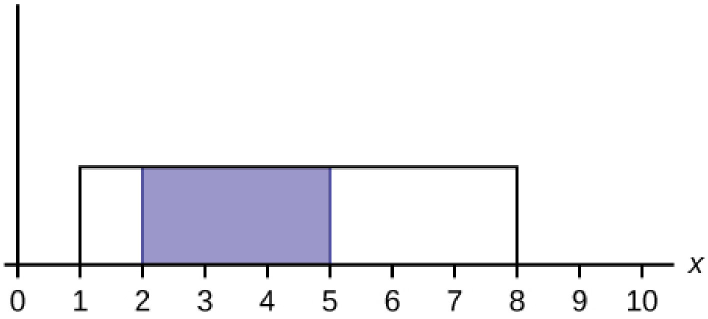
Formula Review

Probability density function (pdf) $f(x)$:

- $f(x) \geq 0$
- The total area under the curve $f(x)$ is one.

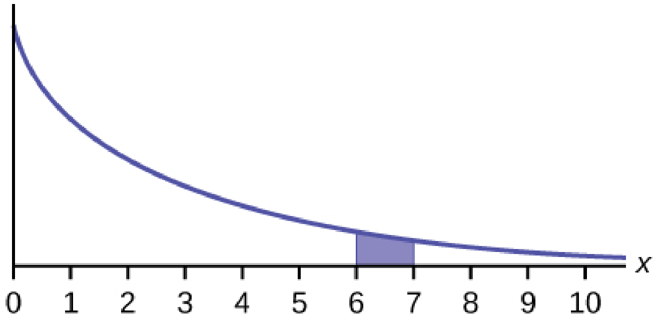
Cumulative distribution function (cdf): $P(X \leq x)$
Exercise:

Problem: What does the shaded area represent? $P($___$< x <$ ___$)$



Exercise:

Problem: What does the shaded area represent? $P($___$< x <$ ___$)$



Solution:

$P(6 < x < 7)$

Exercise:

Problem:

For a continuous probablity distribution, $0 \leq x \leq 15$. What is $P(x > 15)$?

Exercise:

Problem:

What is the area under $f(x)$ if the function is a continuous probability density function?

Solution:

one

Exercise:

Problem:

For a continuous probability distribution, $0 \leq x \leq 10$. What is $P(x = 7)$?

Exercise:

Problem:

A **continuous** probability function is restricted to the portion between $x = 0$ and 7. What is $P(x = 10)$?

Solution:

zero

Exercise:

Problem:

$f(x)$ for a continuous probability function is $\frac{1}{5}$, and the function is restricted to $0 \le x \le 5$. What is $P(x < 0)$?

Exercise:

Problem:

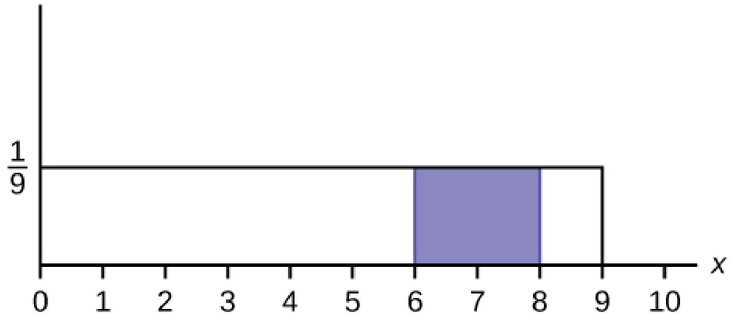
$f(x)$, a continuous probability function, is equal to $\frac{1}{12}$, and the function is restricted to $0 \le x \le 12$. What is $P(0 < x < 12)$?

Solution:

one

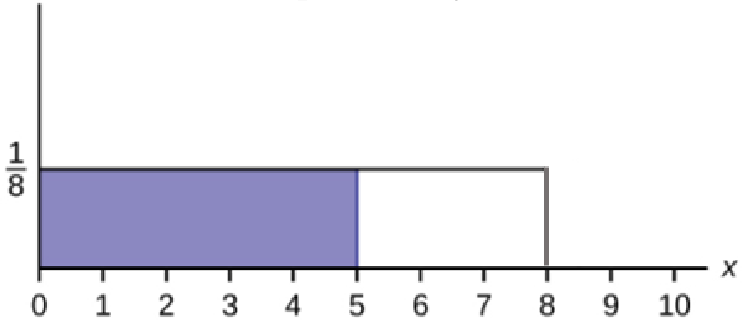
Exercise:

Problem: Find the probability that x falls in the shaded area.



Exercise:

Problem: Find the probability that x falls in the shaded area.

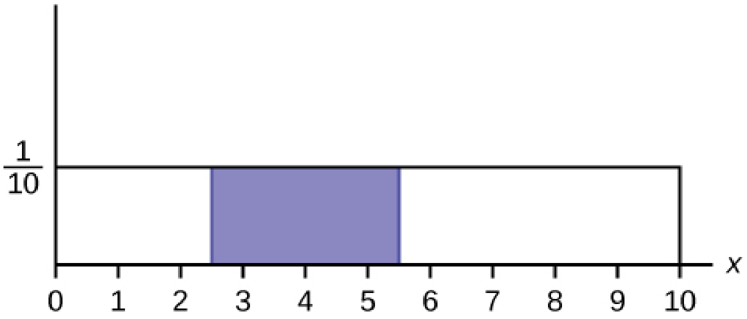


Solution:

0.625

Exercise:

Problem: Find the probability that x falls in the shaded area.



Exercise:

Problem:

$f(x)$, a continuous probability function, is equal to $\frac{1}{3}$ and the function is restricted to $1 \le x \le 4$. Describe $P\left(x > \frac{3}{2}\right)$.

Solution:

The probability is equal to the area from $x = \frac{3}{2}$ to $x = 4$ above the x-axis and up to $f(x) = \frac{1}{3}$.

Homework

Exercise:

Problem:

Consider the following experiment. You are one of 100 people enlisted to take part in a study to determine the percent of nurses in America with an R.N. (registered nurse) degree. You ask nurses if they have an R.N. degree. The nurses answer “yes” or “no.” You then calculate the percentage of nurses with an R.N. degree. You give that percentage to your supervisor.

a. What part of the experiment will yield discrete data?
b. What part of the experiment will yield continuous data?

Exercise:

Problem:

When age is rounded to the nearest year, do the data stay continuous, or do they become discrete? Why?

Solution:

Age is a measurement, regardless of the accuracy used. So the data will stay continuous.

The Uniform Distribution

There are two types of uniform distributions: discrete and continuous. In this section, we will be focusing on the continuous uniform distribution, which is a continuous probability distribution and is concerned with events whose outcomes fall within an interval of real numbers, each of which are equally likely to occur. When working out problems that have a uniform distribution, be careful to note if the data is inclusive or exclusive.

The notation for the uniform distribution is

$X \sim U(a, b)$ where a = the lowest value of x and b = the highest value of x.

The probability density function is $f(x) = \frac{1}{b-a}$ for $a \le x \le b$.

Formulas for the theoretical mean and standard deviation are

$\mu = \frac{a+b}{2}$ and $\sigma = \sqrt{\frac{(b-a)^2}{12}}$

Example:
The data in the following table are 55 smiling times, in seconds, of an eight-week-old baby.

10.4	19.6	18.8	13.9	17.8	16.8	21.6	17.9	12.5	11.1	4.9
12.8	14.8	22.8	20.0	15.9	16.3	13.4	17.1	14.5	19.0	22.8
1.3	0.7	8.9	11.9	10.9	7.3	5.9	3.7	17.9	19.2	9.8
5.8	6.9	2.6	5.8	21.7	11.8	3.4	2.1	4.5	6.3	10.7
8.9	9.4	9.4	7.6	10.0	3.3	6.7	7.8	11.6	13.8	18.6

The sample mean = 11.49 and the sample standard deviation = 6.23.

We will assume that the smiling times, in seconds, follow a uniform distribution between zero and 23 seconds, inclusive. This means that any smiling time from zero to and including 23 seconds is **equally likely**. The histogram that could be constructed from the sample is an empirical distribution that closely matches the theoretical uniform distribution.

Let X = length, in seconds, of an eight-week-old baby's smile.

For this example, $X \sim U(0, 23)$ and $f(x) = \frac{1}{23-0}$ for $0 \le X \le 23$.

For this problem, the theoretical mean and standard deviation are

$\mu = \frac{0+23}{2}$ = 11.50 seconds and $\sigma = \sqrt{\frac{(23-0)^2}{12}}$ = 6.64 seconds.

Notice that the theoretical mean and standard deviation are close to the sample mean and standard deviation in this example.

Example:
Exercise:

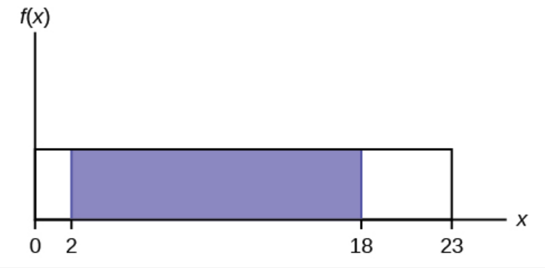
Problem:

a. Refer to [link]. What is the probability that a randomly chosen eight-week-old baby smiles between two and 18 seconds?

Solution:

a. Find $P(2 < x < 18)$.

$P(2 < x < 18)$ = (base)(height) = $(18 - 2)\left(\frac{1}{23}\right) = \frac{16}{23}$.



Exercise:

Problem: b. Find the 90th percentile for an eight-week-old baby's smiling time.

Solution:

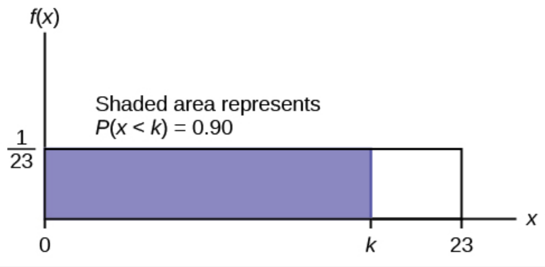
b. Ninety percent of the smiling times fall below the 90th percentile, k, so

$$P(x < k) = 0.90$$

$$(\text{base})(\text{height}) = 0.90$$

$$(k - 0)\left(\frac{1}{23}\right) = 0.90$$

$$k = (23)(0.90) = 20.7$$



Exercise:

Problem:

c. Find the probability that a random eight-week-old baby smiles more than 12 seconds **KNOWING** that the baby smiles **MORE THAN EIGHT SECONDS**.

Solution:

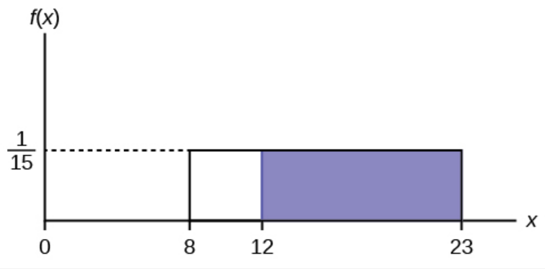
c. This probability question is a **conditional**. You are asked to find the probability that an eight-week-old baby smiles more than 12 seconds when you **already know** the baby has smiled for more than eight seconds.

Find $P(x > 12|x > 8)$ There are two ways to do the problem. **For the first way**, use the fact that this is a **conditional** and changes the sample space. The graph illustrates the new sample space. You already know the baby smiled more than eight seconds.

Write a new $f(x)$: $f(x) = \frac{1}{23-8} = \frac{1}{15}$

for $8 < x < 23$

$P(x > 12|x > 8) = (23 - 12)\left(\frac{1}{15}\right) = \frac{11}{15}$

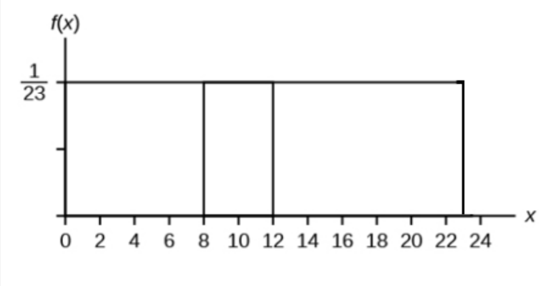


For the second way, use the conditional formula from Probability Topics with the original distribution $X \sim U(0, 23)$:

$P(A|B) = \frac{P(A \text{ AND } B)}{P(B)}$

For this problem, A is $(x > 12)$ and B is $(x > 8)$.

So, $P(x > 12|x > 8) = \frac{P(x>12 \text{ AND } x>8)}{P(x>8)} = \frac{P(x>12)}{P(x>8)} = \frac{\frac{11}{23}}{\frac{15}{23}} = \frac{11}{15}$



Note:
Try It
Exercise:

Problem: A distribution is given as $X \sim U(0, 20)$. What is $P(2 < x < 18)$? Find the 90th percentile.

Solution:

$P(2 < x < 18)$ = 0.8; 90th percentile = 18

Example:
The amount of time, in minutes, that a person must wait for a bus is uniformly distributed between zero and 15 minutes, inclusive.
Exercise:

Problem: a. What is the probability that a person waits fewer than 12.5 minutes?

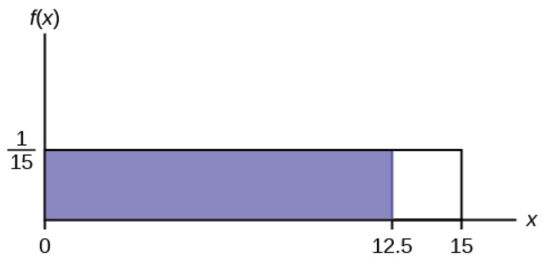
Solution:

a. Let X = the number of minutes a person must wait for a bus. a = 0 and b = 15. $X \sim U(0, 15)$. Write the probability density function. $f(x) = \frac{1}{15-0} = \frac{1}{15}$ for $0 \leq x \leq 15$.

Find $P(x < 12.5)$. Draw a graph.

$P(x < 12.5)$ = (base)(height) = $(12.5 - 0)\left(\frac{1}{15}\right)$ = 0.8333

The probability a person waits less than 12.5 minutes is 0.8333.



Exercise:

Problem:

b. On the average, how long must a person wait? Find the mean, μ, and the standard deviation, σ.

Solution:

b. $\mu = \frac{a+b}{2} = \frac{15+0}{2} = 7.5$. On the average, a person must wait 7.5 minutes.

$\sigma = \sqrt{\frac{(b-a)^2}{12}} = \sqrt{\frac{(15-0)^2}{12}} = 4.3$. The Standard deviation is 4.3 minutes.

Exercise:

Problem: c. Ninety percent of the time, the time a person must wait falls below what value?

Note: This asks for the 90th percentile.

Solution:

c. Find the 90th percentile. Draw a graph. Let k = the 90th percentile.

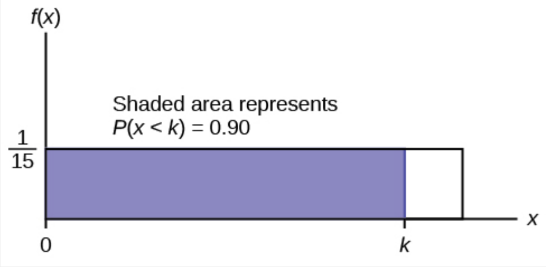
$$P(x < k) = (\text{base})(\text{height}) = (k - 0)(\tfrac{1}{15})$$

$$0.90 = (k)\left(\tfrac{1}{15}\right)$$

$$k = (0.90)(15) = 13.5$$

k is sometimes called a critical value.

The 90th percentile is 13.5 minutes. Ninety percent of the time, a person must wait at most 13.5 minutes.



Note:
Try It
Exercise:

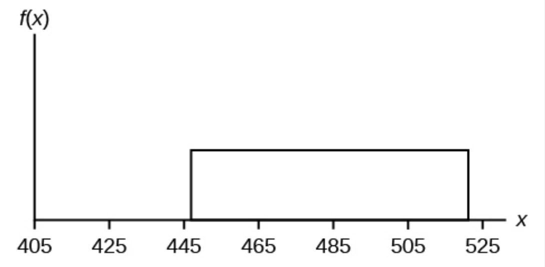
Problem:

The total duration of baseball games in the major league in the 2011 season is uniformly distributed between 447 hours and 521 hours inclusive.

a. Find *a* and *b* and describe what they represent.
b. Write the distribution.
c. Find the mean and the standard deviation.
d. What is the probability that the duration of games for a team for the 2011 season is between 480 and 500 hours?
e. What is the 65th percentile for the duration of games for a team for the 2011 season?

Solution:

a. *a* is 447, and *b* is 521. *a* is the minimum duration of games for a team for the 2011 season, and *b* is the maximum duration of games for a team for the 2011 season.
b. $X \sim U(447, 521)$.
c. $\mu = 484$, and $\sigma = 21.36$



d. $P(480 < x < 500) = 0.2703$
e. 65th percentile is 495.1 hours.

Example:
Suppose the time it takes a nine-year old to eat a donut is between 0.5 and 4 minutes, inclusive. Let X = the time, in minutes, it takes a nine-year old child to eat a donut. Then $X \sim U(0.5, 4)$.
Exercise:

Problem:

a. The probability that a randomly selected nine-year old child eats a donut in at least two minutes is _______.

Solution:

a. 0.5714

Exercise:

Problem:

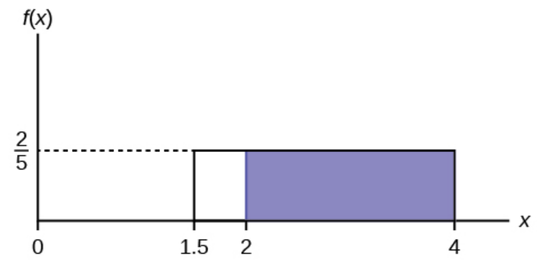
b. Find the probability that a different nine-year old child eats a donut in more than two minutes given that the child has already been eating the donut for more than 1.5 minutes.

The second question has a **conditional probability**. You are asked to find the probability that a nine-year old child eats a donut in more than two minutes given that the child has already been eating the donut for more than 1.5 minutes. Solve the problem two different ways (see [link]). You must reduce the sample space. **First way**: Since you know the child has already been eating the donut for more than 1.5 minutes, you are no longer starting at *a* = 0.5 minutes. Your starting point is 1.5 minutes.

Write a new $f(x)$:

$f(x) = \frac{1}{4-1.5} = \frac{2}{5}$ for $1.5 \le x \le 4$.

Find $P(x > 2|x > 1.5)$. Draw a graph.



$P(x > 2|x > 1.5)$ = (base)(new height) = $(4 - 2)\left(\frac{2}{5}\right)$= ?

Solution:

b. $\frac{4}{5}$

The probability that a nine-year old child eats a donut in more than two minutes given that the child has already been eating the donut for more than 1.5 minutes is $\frac{4}{5}$.

Second way: Draw the original graph for $X \sim U(0.5, 4)$. Use the conditional formula

$$P(x > 2|x > 1.5) = \frac{P(x>2 \text{ AND } x>1.5)}{P(x>1.5)} = \frac{P(x>2)}{P(x>1.5)} = \frac{\frac{2}{3.5}}{\frac{2.5}{3.5}} = 0.8 = \frac{4}{5}$$

Note:
Try It
Exercise:

Problem:

Suppose the time it takes a student to finish a quiz is uniformly distributed between six and 15 minutes, inclusive. Let X = the time, in minutes, it takes a student to finish a quiz. Then $X \sim U(6, 15)$.

Find the probability that a randomly selected student needs at least eight minutes to complete the quiz. Then find the probability that a different student needs at least eight minutes to finish the quiz given that she has already taken more than seven minutes.

Solution:

$P(x > 8) = 0.7778$

$P(x > 8 | x > 7) = 0.875$

Example:
Ace Heating and Air Conditioning Service finds that the amount of time a repairman needs to fix a furnace is uniformly distributed between 1.5 and four hours. Let X = the time needed to fix a furnace. Then $X \sim U(1.5, 4)$.
Exercise:

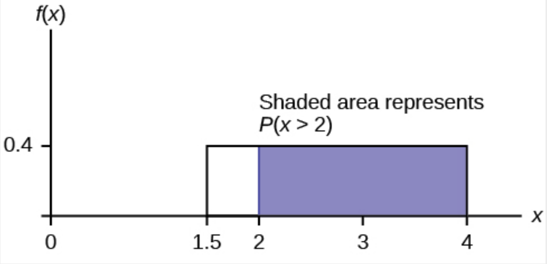
Problem:

a. Find the probability that a randomly selected furnace repair requires more than two hours.
b. Find the probability that a randomly selected furnace repair requires less than three hours.
c. Find the 30th percentile of furnace repair times.
d. The longest 25% of furnace repair times take at least how long? (In other words: find the minimum time for the longest 25% of repair times.) What percentile does this represent?
e. Find the mean and standard deviation

Solution:

a. To find $f(x)$: $f(x) = \frac{1}{4-1.5} = \frac{1}{2.5}$ so $f(x) = 0.4$

$P(x > 2)$ = (base)(height) = (4 – 2)(0.4) = 0.8

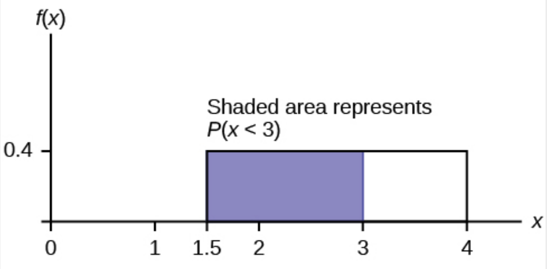


Uniform Distribution between 1.5 and four with shaded area between two and four representing the probability that the repair time *x* is greater than two

Solution:

b. $P(x < 3)$ = (base)(height) = (3 – 1.5)(0.4) = 0.6

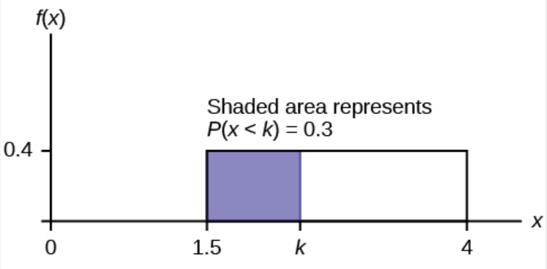
The graph of the rectangle showing the entire distribution would remain the same. However the graph should be shaded between x = 1.5 and x = 3. Note that the shaded area starts at x = 1.5 rather than at x = 0; since $X \sim U(1.5, 4)$, x cannot be less than 1.5.



Uniform Distribution between 1.5 and four with shaded area between 1.5 and three representing the probability that the repair time *x* is less than three

Solution:

c.



Uniform Distribution between 1.5 and 4 with an area of 0.30 shaded to the left, representing the shortest 30% of repair times.

$P(x < k)$ = 0.30

$P(x < k)$ = (base)(height) = $(k – 1.5)(0.4)$

0.3 = $(k – 1.5)(0.4)$; Solve to find k:

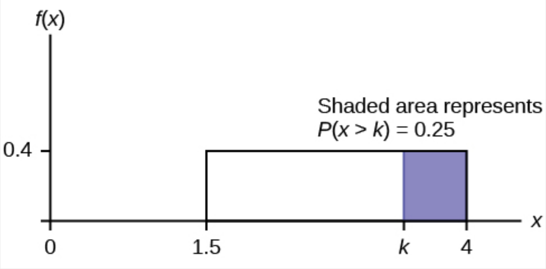
0.75 = k – 1.5, obtained by dividing both sides by 0.4

k = 2.25, obtained by adding 1.5 to both sides

The 30th percentile of repair times is 2.25 hours. 30% of repair times are 2.25 hours or less.

Solution:

d.



Uniform Distribution between 1.5 and 4 with an area of 0.25 shaded to the right representing the longest 25% of repair times.

$P(x < k)$ = 0.25

$P(x < k)$ = (base)(height) = $(4 – k)(0.4)$

0.25 = $(4 – k)(0.4)$; Solve for k:

0.625 = 4 − k, obtained by dividing both sides by 0.4

−3.375 = −k, obtained by subtracting four from both sides: k = 3.375

The longest 25% of furnace repairs take at least 3.375 hours (3.375 hours or longer).

Note: Since 25% of repair times are 3.375 hours or longer, that means that 75% of repair times are 3.375 hours or less. 3.375 hours is the **75th percentile** of furnace repair times.

Solution:

e. $\mu = \frac{a+b}{2}$ and $\sigma = \sqrt{\frac{(b-a)^2}{12}}$

$\mu = \frac{1.5+4}{2} = 2.75$ hours and $\sigma = \sqrt{\frac{(4-1.5)^2}{12}} = 0.7217$ hours

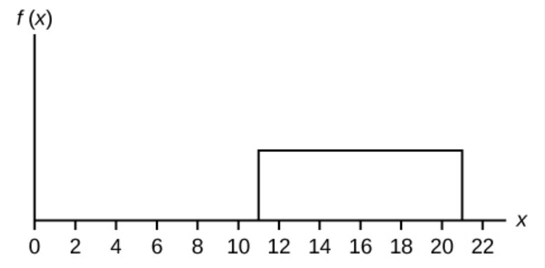
Note:
Try It
Exercise:

Problem:

The amount of time a service technician needs to change the oil in a car is uniformly distributed between 11 and 21 minutes. Let X = the time needed to change the oil on a car.

a. Write the random variable X in words. X = __________________.
b. Write the distribution.
c. Graph the distribution.
d. Find $P(x > 19)$.
e. Find the 50th percentile.

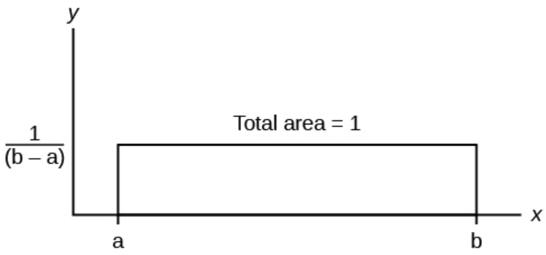
Solution:

a. Let X = the time needed to change the oil in a car.
b. $X \sim U(11, 21)$.
c. 

d. $P(x > 19) = 0.2$
e. the 50th percentile is 16 minutes.

Section Review

If X has a uniform distribution where $a < x < b$ or $a \leq x \leq b$, then X takes on values between a and b (may include a and b). All values x are equally likely. We write $X \sim U(a, b)$. The mean of X is $\mu = \frac{a+b}{2}$. The standard deviation of X is $\sigma = \sqrt{\frac{(b-a)^2}{12}}$. The probability density function of X is $f(x) = \frac{1}{b-a}$ for $a \leq x \leq b$. The cumulative distribution function of X is $P(X \leq x) = \frac{x-a}{b-a}$. X is continuous.



The probability $P(c < X < d)$ may be found by computing the area under $f(x)$, between c and d. Since the corresponding area is a rectangle, the area may be found simply by multiplying the width and the height.

Formula Review

X = a real number between a and b (in some instances, X can take on the values a and b). a = smallest X; b = largest X

$X \sim U(a, b)$

The mean is $\mu = \frac{a+b}{2}$

The standard deviation is $\sigma = \sqrt{\frac{(b-a)^2}{12}}$

Probability density function: $f(x) = \frac{1}{b-a}$ for $a \le X \le b$

Area to the Left of x: $P(X \le x) = (x - a)\left(\frac{1}{b-a}\right)$

Area to the Right of $P(X > x) = (b - x)\left(\frac{1}{b-a}\right)$

Area Between c and d: $P(c < x < d)$ = (base)(height) = $(d - c)\left(\frac{1}{b-a}\right)$

Uniform: $X \sim U(a, b)$ where $a < x < b$

- pdf: $f(x) = \frac{1}{b-a}$ for $a \le x \le b$
- cdf: $P(X \le x) = \frac{x-a}{b-a}$
- mean $\mu = \frac{a+b}{2}$
- standard deviation $\sigma = \sqrt{\frac{(b-a)^2}{12}}$
- $P(c < X < d) = (d - c)\left(\frac{1}{b-a}\right)$

References

McDougall, John A. The McDougall Program for Maximum Weight Loss. Plume, 1995.

Use the following information to answer the next ten questions. The data that follow are the square footage (in 1,000 feet squared) of 28 homes.

1.5	2.4	3.6	2.6	1.6	2.4	2.0
3.5	2.5	1.8	2.4	2.5	3.5	4.0
2.6	1.6	2.2	1.8	3.8	2.5	1.5
2.8	1.8	4.5	1.9	1.9	3.1	1.6

The sample mean = 2.50 and the sample standard deviation = 0.8302.

The distribution can be written as $X \sim U(1.5, 4.5)$.

Exercise:

Problem: What type of distribution is this?

Exercise:

Problem: In this distribution, outcomes are equally likely. What does this mean?

Solution:

It means that the value of x is just as likely to be any number between 1.5 and 4.5.

Exercise:

Problem: What is the height of $f(x)$ for the continuous probability distribution?

Exercise:

Problem: What are the constraints for the values of x?

Solution:

$1.5 \leq x \leq 4.5$

Exercise:

Problem: Graph $P(2 < x < 3)$.

Exercise:

Problem: What is $P(2 < x < 3)$?

Solution:

0.3333

Exercise:

Problem: What is $P(x < 3.5 | x < 4)$?

Exercise:

Problem: What is $P(x = 1.5)$?

Solution:

zero

Exercise:

Problem: What is the 90th percentile of square footage for homes?

Exercise:

Problem:

Find the probability that a randomly selected home has more than 3,000 square feet given that you already know the house has more than 2,000 square feet.

Solution:

0.6

Use the following information to answer the next eight exercises. A distribution is given as $X \sim U(0, 12)$.

Exercise:

Problem: What is a? What does it represent?

Exercise:

Problem: What is b? What does it represent?

Solution:

b is 12, and it represents the highest value of x.

Exercise:

Problem: What is the probability density function?

Exercise:

Problem: What is the theoretical mean?

Solution:

six

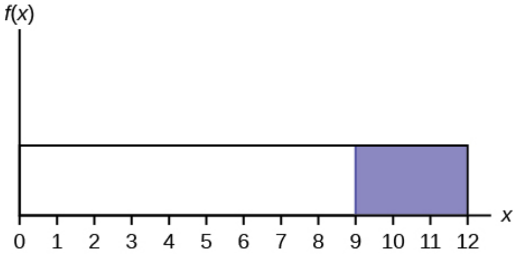
Exercise:

Problem: What is the theoretical standard deviation?

Exercise:

Problem: Draw the graph of the distribution for $P(x > 9)$.

Solution:



Exercise:

Problem: Find $P(x > 9)$.

Exercise:

Problem: Find the 40th percentile.

Solution:

4.8

Use the following information to answer the next eleven exercises. The age of cars in the staff parking lot of a suburban college is uniformly distributed from six months (0.5 years) to 9.5 years.

Exercise:

Problem: What is being measured here?

Exercise:

Problem: In words, define the random variable X.

Solution:

X = The age (in years) of cars in the staff parking lot

Exercise:

Problem: Are the data discrete or continuous?

Exercise:

Problem: The interval of values for x is ______.

Solution:

0.5 to 9.5

Exercise:

Problem: The distribution for X is ______.

Exercise:

Problem: Write the probability density function.

Solution:

$f(x) = \frac{1}{9}$ where x is between 0.5 and 9.5, inclusive.

Exercise:

Problem: Graph the probability distribution.

a. Sketch the graph of the probability distribution.

b. Identify the following values:

i. Lowest value for x: ________
ii. Highest value for x: ________
iii. Height of the rectangle: ________
iv. Label for x-axis (words): ________
v. Label for y-axis (words): ________

Exercise:

Problem: Find the average age of the cars in the lot.

Solution:

$\mu = 5$

Exercise:

Problem: Find the probability that a randomly chosen car in the lot was less than four years old.

a. Sketch the graph, and shade the area of interest.

b. Find the probability. $P(x < 4)$ = _______

Exercise:

Problem:

Considering only the cars less than 7.5 years old, find the probability that a randomly chosen car in the lot was less than four years old.

a. Sketch the graph, shade the area of interest.

b. Find the probability. $P(x < 4 | x < 7.5)$ = _______

Solution:

a. Check student’s solution.
b. $\frac{3.5}{7}$

Exercise:

Problem: What has changed in the previous two problems that made the solutions different?

Exercise:

Problem:

Find the third quartile of ages of cars in the lot. This means you will have to find the value such that $\frac{3}{4}$, or 75%, of the cars are at most (less than or equal to) that age.

a. Sketch the graph, and shade the area of interest.

b. Find the value k such that $P(x < k) = 0.75$.
c. The third quartile is _______

Solution:

a. Check student's solution.
b. $k = 7.25$
c. 7.25

Homework

For each probability and percentile problem, draw the picture.

Exercise:

Problem:

Births are approximately uniformly distributed between the 52 weeks of the year. They can be said to follow a uniform distribution from one to 53 (spread of 52 weeks).

a. $X \sim$ _________
b. Graph the probability distribution.
c. $f(x) =$ _________
d. $\mu =$ _________
e. $\sigma =$ _________
f. Find the probability that a person is born at the exact moment week 19 starts. That is, find $P(x = 19) =$ _________
g. $P(2 < x < 31) =$ _________
h. Find the probability that a person is born after week 40.
i. $P(12 < x | x < 28) =$ _________
j. Find the 70th percentile.
k. Find the value for the third quartile.

Exercise:

Problem: A random number generator picks a number from one to nine in a uniform manner.

a. $X \sim$ _________
b. Graph the probability distribution.
c. $f(x) =$ _________
d. $\mu =$ _________
e. $\sigma =$ _________

f. $P(3.5 < x < 7.25)$ = __________
g. $P(x > 5.67)$
h. $P(x > 5|x > 3)$ = __________
i. Find the 90th percentile.

Solution:

a. $X \sim U(1, 9)$
b. Check student's solution.
c. $f(x) = \frac{1}{8}$ where $1 \leq x \leq 9$
d. five
e. 2.3
f. $\frac{15}{32}$
g. $\frac{333}{800}$
h. $\frac{2}{3}$
i. 8.2

Exercise:

Problem:

According to a study by Dr. John McDougall of his live-in weight loss program at St. Helena Hospital, the people who follow his program lose between six and 15 pounds a month until they approach trim body weight. Let's suppose that the weight loss is uniformly distributed. We are interested in the weight loss of a randomly selected individual following the program for one month.

a. Define the random variable. X = __________
b. $X \sim$ __________
c. Graph the probability distribution.
d. $f(x)$ = __________
e. μ = __________
f. σ = __________
g. Find the probability that the individual lost more than ten pounds in a month.
h. Suppose it is known that the individual lost more than ten pounds in a month. Find the probability that he lost less than 12 pounds in the month.
i. $P(7 < x < 13|x > 9)$ = ___________. State this in a probability question, similarly to parts g and h, draw the picture, and find the probability.

Exercise:

Problem:

A subway train on the Red Line arrives every eight minutes during rush hour. We are interested in the length of time a commuter must wait for a train to arrive. The time follows a uniform distribution.

a. Define the random variable. X = __________
b. $X \sim$ __________
c. Graph the probability distribution.
d. $f(x)$ = __________
e. μ = __________
f. σ = __________

g. Find the probability that the commuter waits less than one minute.
h. Find the probability that the commuter waits between three and four minutes.
i. Sixty percent of commuters wait more than how long for the train? State this in a probability question, similarly to parts g and h, draw the picture, and find the probability.

Solution:

a. X represents the length of time a commuter must wait for a train to arrive on the Red Line.
b. X ~ $U(0, 8)$
c. Check student's solution.
d. $f(x) = \frac{1}{8}$ where ≤ x ≤ 8
e. four
f. 2.31
g. $\frac{1}{8}$
h. $\frac{1}{8}$
i. 3.2

Exercise:

Problem:

The age of a first grader on September 1 at Garden Elementary School is uniformly distributed from 5.8 to 6.8 years. We randomly select one first grader from the class.

a. Define the random variable. X = _________
b. X ~ _________
c. Graph the probability distribution.
d. $f(x)$ = _________
e. μ = _________
f. σ = _________
g. Find the probability that she is over 6.5 years old.
h. Find the probability that she is between four and six years old.
i. Find the 70th percentile for the age of first graders on September 1 at Garden Elementary School.

Use the following information to answer the next three exercises. The Sky Train from the terminal to the rental–car and long–term parking center is supposed to arrive every eight minutes. The waiting times for the train are known to follow a uniform distribution.

Exercise:

Problem: What is the average waiting time (in minutes)?

a. zero
b. two
c. three
d. four

Solution:

d

Exercise:

Problem: Find the 30th percentile for the waiting times (in minutes).

a. two
b. 2.4
c. 2.75
d. three

Exercise:

Problem:

The probability of waiting more than seven minutes given a person has waited more than four minutes is?

a. 0.125
b. 0.25
c. 0.5
d. 0.75

Solution:

b

Exercise:

Problem:

The time (in minutes) until the next bus departs a major bus depot follows a distribution with $f(x) = \frac{1}{20}$ where x goes from 25 to 45 minutes.

a. Define the random variable. X = _________
b. X ~ _________
c. Graph the probability distribution.
d. $f(x)$ = _________
e. μ = _________
f. σ = _________
g. Find the probability that the time is at most 30 minutes. Sketch and label a graph of the distribution. Shade the area of interest. Write the answer in a probability statement.
h. Find the probability that the time is between 30 and 40 minutes. Sketch and label a graph of the distribution. Shade the area of interest. Write the answer in a probability statement.
i. $P(25 < x < 55)$ = _________. State this in a probability statement, similarly to parts g and h, draw the picture, and find the probability.
j. Find the 90th percentile. This means that 90% of the time, the time is less than _____ minutes.
k. Find the 75th percentile. In a complete sentence, state what this means. (See part j.)
l. Find the probability that the time is more than 40 minutes given (or knowing that) it is at least 30 minutes.

Exercise:

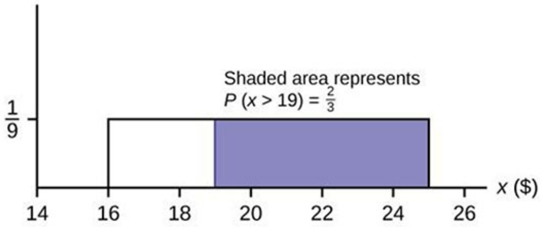
Problem:

Suppose that the value of a stock varies each day from \$16 to \$25 with a uniform distribution.

a. Find the probability that the value of the stock is more than \$19.
b. Find the probability that the value of the stock is between \$19 and \$22.
c. Find the upper quartile - 25% of all days the stock is above what value?
d. Given that the stock is greater than \$18, find the probability that the stock is more than \$21.

Solution:

a. The probability density function of X is $\frac{1}{25-16} = \frac{1}{9}$.
   $P(X > 19) = (25 – 19) \left(\frac{1}{9}\right) = \frac{6}{9} = \frac{2}{3}$.



b. $P(19 < x < 22) = (22 – 19) \left(\frac{1}{9}\right) = \frac{3}{9} = \frac{1}{3}$.
c. The area must be 0.25, and 0.25 = (width)$\left(\frac{1}{9}\right)$, so width = (0.25)(9) = 2.25. Thus, the value is 25 – 2.25 = 22.75.
d. This is a conditional probability question. $P(x > 21|x > 18)$. You can do this two ways:

   - Draw the graph where a is now 18 and b is still 25. The height is $\frac{1}{(25-18)} = \frac{1}{7}$
     So, $P(x > 21|x > 18) = (25 – 21)\left(\frac{1}{7}\right) = 4/7$.
   - Use the formula: $P(x > 21|x > 18) = \frac{P(x>21 \text{ AND } x>18)}{P(x>18)}$
     $= \frac{P(x>21)}{P(x>18)} = \frac{(25-21)}{(25-18)} = \frac{4}{7}$.

Exercise:

Problem:

A fireworks show is designed so that the time between fireworks is between one and five seconds, and follows a uniform distribution.

a. Find the average time between fireworks.
b. Find probability that the time between fireworks is greater than four seconds.

Exercise:

Problem:

The distance, in miles, driven by a truck driver falls between 300 and 700, and follows a uniform distribution.

a. Find the probability that the truck driver goes more than 650 miles in a day.
b. Find the probability that the truck drivers goes between 400 and 650 miles in a day.

c. At least how many miles does the truck driver travel on the furthest 10% of days?

Solution:

a. $P(x > 650) = \frac{700-650}{700-300} = \frac{50}{400} = \frac{1}{8} = 0.125$.

b. $P(400 < X < 650) = \frac{650-400}{700-300} = \frac{250}{400} = 0.625$

c. $0.10 = \frac{\text{width}}{700-300}$, so width = 400(0.10) = 40. Since 700 – 40 = 660, the drivers travel at least 660 miles on the furthest 10% of days.

Glossary

Conditional Probability
: the likelihood that an event will occur given that another event has already occurred.

The Exponential Distribution (Optional)

The **exponential distribution** is often concerned with the amount of time until some specific event occurs. For example, the amount of time (beginning now) until an earthquake occurs has an exponential distribution. Other examples include the length, in minutes, of long distance business telephone calls, and the amount of time, in months, a car battery lasts.

Values for an exponential random variable occur in the following way. There are fewer large values and more small values. For example, the amount of money customers spend in one trip to the supermarket follows an exponential distribution. There are more people who spend small amounts of money and fewer people who spend large amounts of money.

The exponential distribution is widely used in the field of reliability. Reliability deals with the amount of time a product lasts.

If X is a random variable and has an **exponential distribution** with mean μ, then the **decay parameter**, m is given by the formula $m = \frac{1}{\mu}$, and we write X ∼ *Exp*(m) where $x \geq 0$ and $m > 0$.

The standard deviation, σ, is the same as the mean; $\mu = \sigma$.

The probability density function of X is $f\left(x\right) = me^{-mx}$ (or equivalently $f(x) = \frac{1}{\mu}e^{-x/\mu}$).

The cumulative distribution function of X is $P\left(X \leq x\right) = 1 - e^{-mx}$.

Example:
Let X = amount of time (in minutes) a postal clerk spends with his or her customer. The time is known to have an exponential distribution with the average amount of time equal to four minutes.

X is a **continuous random variable** since time is measured. It is given that μ = 4 minutes. To do any calculations, you must know m, the decay parameter.

$m = \frac{1}{\mu}$. Therefore, $m = \frac{1}{4} = 0.25$.

The distribution notation is X ~ *Exp*(m). Therefore, X ~ *Exp*(0.25).

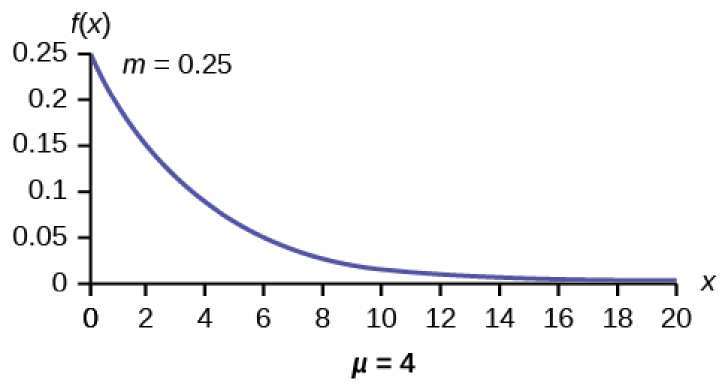
The probability density function is $f\left(x\right) = me^{-mx}$. The number *e* = 2.71828182846... It is a number that is used often in mathematics. Scientific calculators have the key "e^x." If you enter one for x, the calculator will display the value e.

The curve is:

$f\left(x\right) = 0.25e^{-0.25x}$ where x is at least zero and m = 0.25.

For example, $f\left(5\right) = 0.25e^{-0.25(5)} = 0.072$.

The graph is as follows:



Notice the graph is a declining curve. When x = 0,

$f\left(0\right) = 0.25e^{-0.25(0)} = \left(0.25\right)\left(1\right) = 0.25 = m$. The maximum value on the y-axis is m.

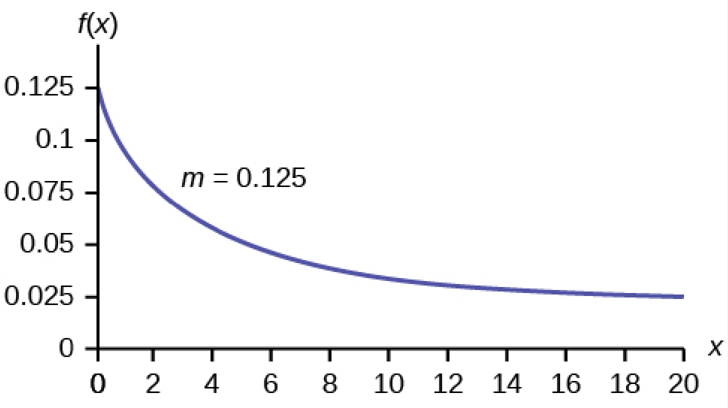
Note:
Try It
Exercise:

Problem:

The amount of time spouses shop for anniversary cards can be modeled by an exponential distribution with the average amount of time equal to eight minutes. Write the distribution, state the probability density function, and graph the distribution.

Solution:

X ~ *Exp*(0.125); $f\left(x\right)=0.125{e}^{-0.125x}$.



Example:
Exercise:

Problem:

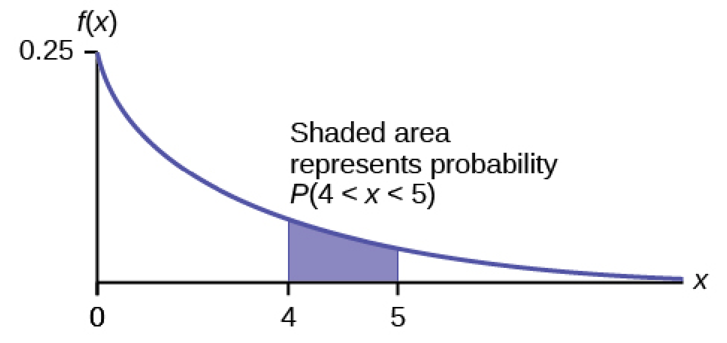
a. Using the information in [link], find the probability that a clerk spends four to five minutes with a randomly selected customer.

Solution:

a. Find $P(4<x<5)$.

The **cumulative distribution function (CDF)** gives the area to the left: $P\left(X<x\right)=1-{e}^{-mx}$, where $m=0.25$.

So, $P\left(x < 5\right) = 1 - e^{-0.25(5)} = 0.7135$ and $P\left(x < 4\right) = 1 - e^{-0.25(4)} = 0.6321$.



Note:
NOTE
You can do these calculations easily on a calculator.

The probability that a postal clerk spends four to five minutes with a randomly selected customer is $P\left(4 < x < 5\right) = P\left(x < 5\right) - P\left(x < 4\right) = 0.7135 - 0.6321 = 0.0814$.

Note:
On the home screen, enter (1 – e^(–0.25*5))–(1–e^(–0.25*4)) or simplify and enter e^(–0.25*4) – e^(–0.25*5).

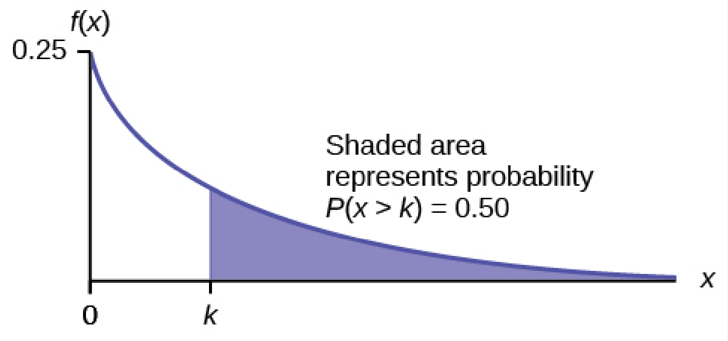
Exercise:

Problem:

b. Half of all customers are finished within how long? (Find the 50th percentile)

Solution:

b. Find the 50th percentile.



Find k such that $P\left(x < k\right) = 0.5$.

Note that $P\left(x < k\right) = 0.5$ and $P\left(x < k\right) = 1 - e^{-0.25k}$.

Therefore, $0.5 = 1 - e^{-0.25k}$, which means $e^{-0.25k} = 0.5$.

Take the natural log of both sides: $\ln\left(e^{-0.25k}\right) = \ln\left(0.5\right)$. So, $-0.25k = \ln\left(0.5\right)$.

Solve for k to obtain $k = \frac{\ln(0.50)}{-0.25} = 2.8$ minutes. Half of all customers are finished within 2.8 minutes.

Note:
Note
A formula for the k^{th} percentile is: $k = \frac{\ln(1-\text{Area to the left of k})}{-m}$ where ln is the natural log.

Exercise:

Problem:

c. Which is larger, the mean or the median?

Solution:

c. From part b, the median or 50^{th} percentile is 2.8 minutes. The theoretical mean is four minutes. The mean is larger.

Note:
Try It
Exercise:

Problem:

The number of days ahead travelers purchase their airline tickets can be modeled by an exponential distribution with the average amount of time equal to 15 days. Find the probability that a traveler will purchase a ticket fewer than ten days in advance. How many days do half of all travelers wait?

Solution:

$P(x < 10) = 0.4866$

50^{th} percentile = 10.40

Note:
Collaborative Exercise
Have each class member count the change he or she has in his or her pocket or purse. Your instructor will record the amounts in dollars and cents. Construct a histogram of the data taken by the class. Use five intervals. Draw a smooth curve through the bars. The graph should look approximately exponential. Then calculate the mean.

Let X = the amount of money a student in your class has in his or her pocket or purse.
The distribution for X is approximately exponential with mean, μ = _______ and m = _______. The standard deviation, σ = ________.
Draw the appropriate exponential graph. You should label the x– and y–axes, the decay rate, and the mean. Shade the area that represents the probability that one student has less than \$0.40 in his or her pocket or purse. (Shade $P(x < 0.40)$).

Example:
On the average, a certain computer part lasts ten years. The length of time the computer part lasts is exponentially distributed.
Exercise:

Problem:

a. What is the probability that a computer part lasts more than 7 years?

Solution:

a. Let X = the amount of time (in years) a computer part lasts.

$\mu = 10$ so $m = \frac{1}{\mu} = \frac{1}{10} = 0.1$

Find $P\left(x > 7\right)$.

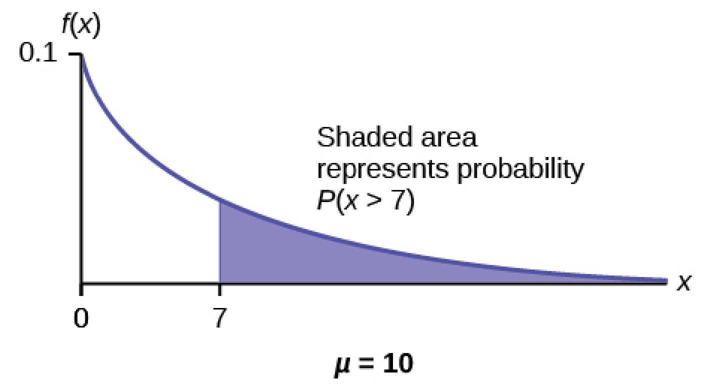
$P\left(x > 7\right) = 1 - P\left(x < 7\right)$.

Since $P\left(X < x\right) = 1 - e^{-mx}$ it follows that
$P\left(X > x\right) = 1 - \left(1 - e^{-mx}\right) = e^{-mx}$.

Therefore, $P\left(x > 7\right) = e^{-0.1(7)} = 0.4966$.

The probability that a computer part lasts more than seven years is 0.4966.

Note:
On the home screen, enter e^(-.1*7).

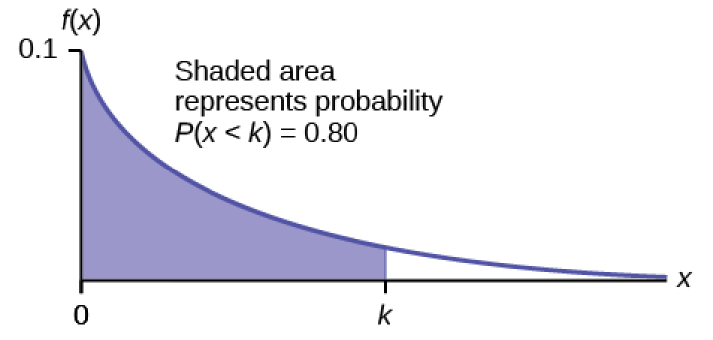


Exercise:

Problem: c. Eighty percent of computer parts last at most how long?

Solution:

c. Find the 80th percentile. Draw the graph. Let k = the 80th percentile.



Solve for k: $k = \frac{ln(1–0.80)}{–0.1} = 16.1$ years

Eighty percent of the computer parts last at most 16.1 years.

Note:

On the home screen, enter ln(1 - .8)/-.1

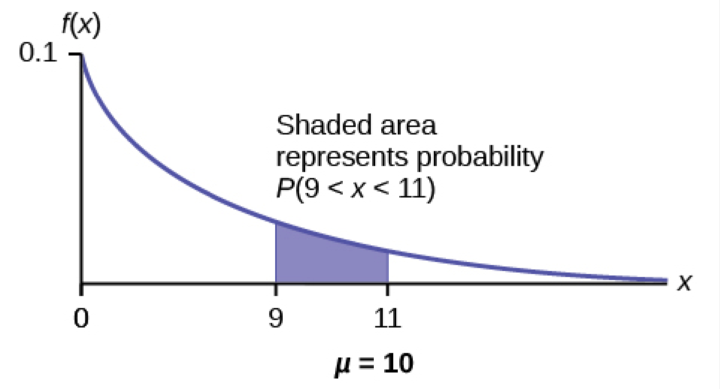
Exercise:

Problem:

d. What is the probability that a computer part lasts between nine and 11 years?

Solution:

d. Find $P\left(9 < x < 11\right)$. Draw the graph.



$$P\left(9 < x < 11\right) = P\left(x < 11\right) - P\left(x < 9\right) = \left(1 - e^{-0.1(11)}\right) - \left(1 - e^{-0.1(9)}\right) = 0.6671 - 0.5934 = 0.0737$$

The probability that a computer part lasts between nine and 11 years is 0.0737.

Note:
On the home screen, after simplifying, enter *e*^(–0.1*9) – *e*^(–0.1*11).

Note:
Try It
Exercise:

Problem:

On average, a pair of running shoes can last 18 months if used every day. The length of time running shoes last is exponentially distributed. What is the probability that a pair of running shoes last more than 15 months? Eighty percent of running shoes last at most how long if used every day? (Note: Use the exact value for m.)

Solution:

$P\left(x > 15\right) = 0.4346$

80th percentile = 28.97 months

Example:
Suppose that the length of a phone call, in minutes, is an exponential random variable with decay parameter $m = \frac{1}{12}$. If another person arrives at a public telephone just before you, find the probability that you will have to wait more than five minutes. Let X = the length of a phone call, in minutes.
Exercise:

Problem:

What is m, μ, and σ? The probability that you must wait more than five minutes is ________ .

Solution:

- $m = \frac{1}{12}$
- $\mu = 12$
- $\sigma = 12$

$P\left(x > 5\right) = 0.6592$

Note:
Try It
Exercise:

Problem:

Suppose that the distance, in miles, that people are willing to commute to work is an exponential random variable with a decay parameter $\frac{1}{20}$. Let X = the distance people are willing to commute in miles. What is m, μ, and σ? What is the probability that a person is willing to commute more than 25 miles?

Solution:

$m = \frac{1}{20};\ \mu = 20;\ \sigma = 20;\ P\left(x > 25\right) = 0.2865$

Example:
The time spent waiting between events is often modeled using the exponential distribution. For example, suppose that an average of 30 customers per hour arrive at a store and the time between arrivals is exponentially distributed.
Exercise:

Problem:

a. On average, how many minutes elapse between two successive arrivals?
b. When the store first opens, how long on average does it take for three customers to arrive?
c. After a customer arrives, find the probability that it takes less than one minute for the next customer to arrive.
d. After a customer arrives, find the probability that it takes more than five minutes for the next customer to arrive.
e. Seventy percent of the customers arrive within how many minutes of the previous customer?
f. Is an exponential distribution reasonable for this situation?

Solution:

a. Since we expect 30 customers to arrive per hour (60 minutes), we expect on average one customer to arrive every two minutes on average.

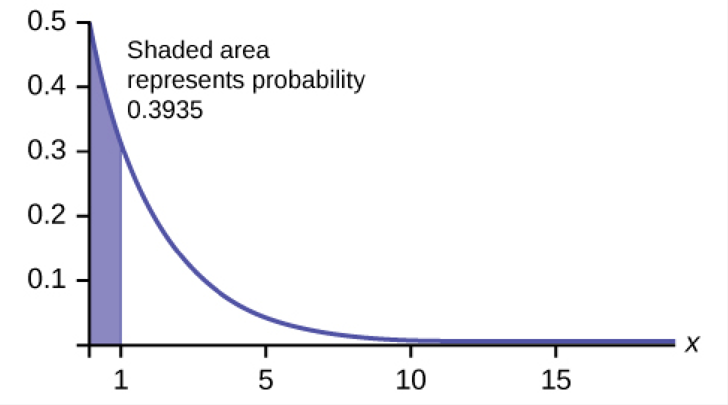
b. Since one customer arrives every two minutes on average, it will take six minutes on average for three customers to arrive.

c. Let X = the time between arrivals, in minutes. By part a, $\mu = 2$, so $m = \frac{1}{2} = 0.5$. Therefore, $X \sim$ *Exp*(0.5).

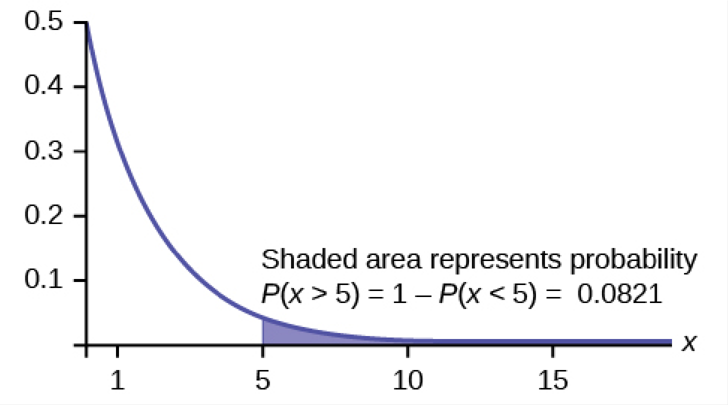
The cumulative distribution function is $P\left(X < x\right) = 1 - e^{-0.5x}$.

Therefore, $P\left(x < 1\right) = 1 - e^{-0.5(1)} = 0.3935$.

Note:
1 - *e*^(–0.5) = 0.3935



d. $P\left(x > 5\right) = 1 - P\left(x < 5\right) = 1 - \left(1 - e^{-0.5(5)}\right) = e^{-2.5} = 0.0821$.

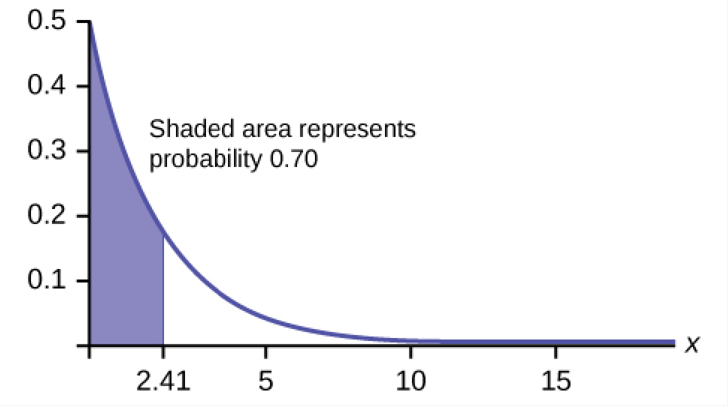


Note:
1 – (1 – e^(– 5*0.5)) or e^(– 5*0.5) = 0.0821

e. This is equivalent to finding the 70th percentile.

   So, $k = \frac{\ln(1-0.70)}{-0.5}$. This gives us $k = 2.41$ minutes.

   Thus, seventy percent of customers arrive within 2.41 minutes of the previous customer.



f. This model assumes that a single customer arrives at a time, which may not be reasonable since people might shop in groups, leading to several customers arriving at the same time. It also assumes that the flow of customers does not change throughout the day, which is not valid if some times of the day are busier than others.

Note:
Try It
Exercise:

Problem:

Suppose that on a certain stretch of highway, cars pass at an average rate of five cars per minute. Assume that the duration of time between successive cars follows the exponential distribution.

a. On average, how many seconds elapse between two successive cars?
b. After a car passes by, how long on average will it take for another seven cars to pass by?
c. Find the probability that after a car passes by, the next car will pass within the next 20 seconds.
d. Find the probability that after a car passes by, the next car will not pass for at least another 15 seconds.

Solution:

a. At a rate of five cars per minute, we expect $\frac{60}{5}$ = 12 seconds to pass between successive cars on average.
b. Using the answer from part a, we see that it takes (12)(7) = 84 seconds for the next seven cars to pass by.
c. Let T = the time (in seconds) between successive cars.
   The mean of T is 12 seconds, so the decay parameter is $\frac{1}{12}$ and $T \sim$ *Exp*($\frac{1}{12}$).

   The cumulative distribution function of T is $P\left(T < t\right) = 1 - e^{-\frac{1}{12}t}$.

   So $P\left(t < 20\right) = 1 - e^{-\frac{1}{12}(20)} = 0.8111$.
d.
   $P\left(t > 15\right) = 1 - P\left(t < 15\right) = 1 - \left(1 - e^{-\frac{1}{12}(15)}\right) = e^{-\frac{15}{12}} = 0.2865$
   .

Memorylessness of the Exponential Distribution

In [link] recall that the amount of time between customers is exponentially distributed with a mean of two minutes (X ~ *Exp* (0.5)). Suppose that five minutes have elapsed since the last customer arrived. Since an unusually long amount of time has now elapsed, it would seem to be more likely for a customer to arrive within the next minute. With the exponential distribution, this is not the case–the additional time spent waiting for the next customer does not depend on how much time has already elapsed since the last customer. This is referred to as the **memoryless property**. Specifically, the **memoryless property** says that

$$P\left(X>r+t|X>r\right)=P\left(X>t\right)$$ for all $r \geq 0$ and $t \geq 0$.

For example, if five minutes has elapsed since the last customer arrived, then the probability that more than one minute will elapse before the next customer arrives is computed by using $r = 5$ and $t = 1$ in the foregoing equation.

$$P\left(x>5+1|x>5\right)=P\left(x>1\right)={e}^{-0.5\left(1\right)}=0.6065.$$

This is the same probability as that of waiting more than one minute for a customer to arrive after the previous arrival.

The exponential distribution is often used to model the longevity of an electrical or mechanical device. In [link], the lifetime of a certain computer part has the exponential distribution with a mean of ten years (X ~ *Exp*(0.1)). The **memoryless property** says that knowledge of what has occurred in the past has no effect on future probabilities. In this case it means that an old part is not any more likely to break down at any particular time than a brand new part. In other words, the part stays as good as new until it suddenly breaks. For example, if the part has already lasted ten years, then the probability that it lasts another seven years is $P\left(x>17|x>10\right)=P\left(x>7\right)=0.4966.$

Example:

Refer to [link] where the time a postal clerk spends with his or her customer has an exponential distribution with a mean of four minutes. Suppose a customer has spent four minutes with a postal clerk. What is the probability that he or she will spend at least an additional three minutes with the postal clerk?

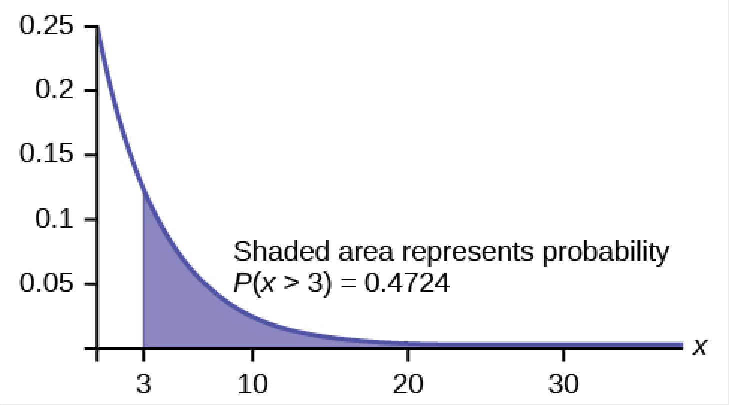
The decay parameter of X is $m = \frac{1}{4}$ = 0.25, so $X \sim Exp(0.25)$.

The cumulative distribution function is $P\left(X < x\right) = 1 - e^{-0.25x}$.

We want to find $P\left(x > 7|x > 4\right)$.

The **memoryless property** says that $P\left(x > 7|x > 4\right) = P\left(x > 3\right)$, so we just need to find the probability that a customer spends more than three minutes with a postal clerk.

This is
$P\left(x > 3\right) = 1 - P\left(x < 3\right) = 1 - \left(1 - e^{-0.25(3)}\right) = e^{-0.75} = 0.4724.$



Note:

1–(1–e^(–0.25*3)) = e^(–0.25*3).

Note:
Try It
Exercise:

Problem:

Suppose that the longevity of a light bulb is exponential with a mean lifetime of eight years. If a bulb has already lasted 12 years, find the probability that it will last a total of over 19 years.

Solution:

Let T = the lifetime of the light bulb. Then T ~ *Exp*$\left(\frac{1}{8}\right)$.

The cumulative distribution function is $P\left(T < t\right) = 1 - e^{-\frac{1}{8}t}$.

We need to find $P\left(t > 19|t = 12\right)$. By the **memoryless property**,

$P\left(t > 19|t = 12\right) = P\left(t > 7\right) = 1 - \left(1 - e^{-\frac{1}{8}(7)}\right) = e^{-\frac{7}{8}} = 0.4169$.

Note:
1 – (1 – e^(–7/8)) = e^(–7/8).

Chapter Review

If X is a random variable and has an **exponential distribution** with mean μ, then the **decay parameter**, m is given by the formula $m = \frac{1}{\mu}$, and we write X ~ *Exp*(m) where $x \geq 0$ and $m > 0$.

The standard deviation, σ, is the same as the mean; $\mu = \sigma$.

The probability density function of X is $f\left(x\right) = me^{-mx}$ (or equivalently

$f(x) = \frac{1}{\mu} e^{-x/\mu}$).

The cumulative distribution function of X is $P\left(X \le x\right) = 1 - e^{-mx}$.

A formula for the percentile k of an exponential distribution is: $k = \frac{\ln(1 - \text{Area to the left of k})}{-m}$ where $\ln$ is the natural log.

The exponential distribution has the **memoryless property**, which says that future probabilities do not depend on any past information. Mathematically, it says that $P\left(X > r + t | X > r\right) = P\left(X > t\right)$ for all r ≥ 0 and t ≥ 0.

Formula Review

Exponential: X ~ *Exp*(m) where m = the decay parameter

- pdf: $f\left(x\right) = me^{-mx}$ where x ≥ 0 and m > 0
- cdf: $P\left(X \le x\right) = 1 - e^{-mx}$
- mean $\mu = \frac{1}{m}$
- standard deviation $\sigma = \mu$
- k^{th} percentile: $k = \frac{\ln(1 - \text{Area to the left of k})}{-m}$
- Additionally
    - $P\left(X > x\right) = e^{-mx}$
    - $P\left(a < X < b\right) = e^{-ma} - e^{-mb}$
- Memoryless Property: $P\left(X > r + t | X > r\right) = P\left(X > t\right)$ for all r ≥ 0 and t ≥ 0

References

Data from the United States Census Bureau.

Data from World Earthquakes, 2013. Available online at http://www.world-earthquakes.com/ (accessed June 11, 2013).

“No-hitter.” Baseball-Reference.com, 2013. Available online at http://www.baseball-reference.com/bullpen/No-hitter (accessed June 11, 2013).

Zhou, Rick. “Exponential Distribution lecture slides.” Available online at www.public.iastate.edu/~riczw/stat330s11/lecture/lec13.pdf (accessed June 11, 2013).

Use the following information to answer the next ten exercises. A customer service representative must spend different amounts of time with each customer to resolve various concerns. The amount of time spent with each customer can be modeled by the following distribution: $X \sim Exp(0.2)$

Exercise:

Problem: What type of distribution is this?

Exercise:

Problem:

Are outcomes equally likely in this distribution? Why or why not?

Solution:

No, outcomes are not equally likely. In this distribution, more people require a little bit of time, and fewer people require a lot of time, so it is more likely that someone will require less time.

Exercise:

Problem: What is m? What does it represent?

Exercise:

Problem: What is the mean?

Solution:

five

Exercise:

Problem: What is the standard deviation?

Exercise:

Problem: State the probability density function.

Solution:

$f\left(x\right)=0.2e^{-0.2x}$

Exercise:

Problem: Graph the distribution.

Exercise:

Problem: Find $P\left(2<x<10\right)$.

Solution:

0.5350

Exercise:

Problem: Find $P\left(x>6\right)$.

Exercise:

Problem: Find the 70th percentile.

Solution:

6.02

Use the following information to answer the next seven exercises. A distribution is given as X ~ *Exp*(0.75).

Exercise:

Problem: What is m?

Exercise:

Problem: What is the probability density function?

Solution:

$f\left(x\right)=0.75{e}^{-0.75x}$

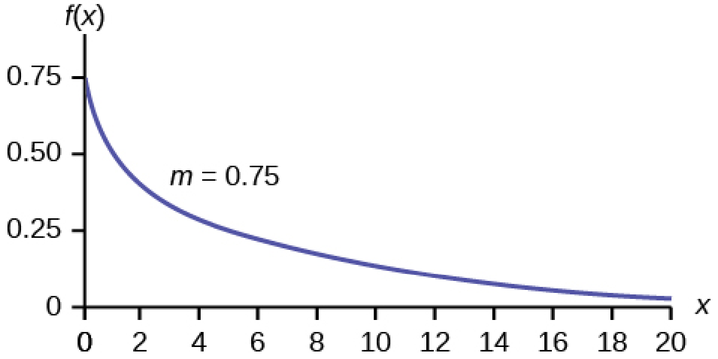
Exercise:

Problem: What is the cumulative distribution function?

Exercise:

Problem: Draw the distribution.

Solution:



Exercise:

Problem: Find $P\left(x<4\right)$.

Exercise:

Problem: Find the 30th percentile.

Solution:

0.4756

Exercise:

Problem: Find the median.

Exercise:

Problem: Which is larger, the mean or the median?

Solution:

The mean is larger. The mean is $\frac{1}{m} = \frac{1}{0.75} \approx 1.33$, which is greater than 0.9242.

Use the following information to answer the next 16 exercises. Carbon-14 is a radioactive element with a half-life of about 5,730 years. Carbon-14 is said to decay exponentially. The decay rate is 0.000121. We start with one gram of carbon-14. We are interested in the time (years) it takes to decay carbon-14.

Exercise:

Problem: What is being measured here?

Exercise:

Problem: Are the data discrete or continuous?

Solution:

continuous

Exercise:

Problem: In words, define the random variable X.

Exercise:

Problem: What is the decay rate (m)?

Solution:

m = 0.000121

Exercise:

Problem: The distribution for X is ______.

Exercise:

Problem:

Find the amount (percent of one gram) of carbon-14 lasting less than 5,730 years. This means, find $P\left(x < 5730\right)$.

a. Sketch the graph, and shade the area of interest.

b. Find the probability. $P\left(x < 5730\right)$ = __________

Solution:

a. Check student's solution
b. $P\left(x < 5730\right) = 0.5001$

Exercise:

Problem:

Find the percentage of carbon-14 lasting longer than 10,000 years.

a. Sketch the graph, and shade the area of interest.

b. Find the probability. $P\left(x > 10,000\right)$ = ________

Exercise:

Problem:

Thirty percent (30%) of carbon-14 will decay within how many years?

a. Sketch the graph, and shade the area of interest.

b. Find the value k such that $P\left(x < k\right) = 0.30$.

Solution:

a. Check student's solution.
b. k = 2947.73

Homework

Exercise:

Problem:

Suppose that the length of long distance phone calls, measured in minutes, is known to have an exponential distribution with the average length of a call equal to eight minutes.

a. Define the random variable. X = _________________.
b. Is X continuous or discrete?
c. X ~ ________
d. μ = ________
e. σ = ________
f. Draw a graph of the probability distribution. Label the axes.
g. Find the probability that a phone call lasts less than nine minutes.
h. Find the probability that a phone call lasts more than nine minutes.
i. Find the probability that a phone call lasts between seven and nine minutes.

Exercise:

Problem:

Suppose that the useful life of a particular car battery, measured in months, decays with parameter 0.025. We are interested in the life of the battery.

a. Define the random variable. X = ___________________________________.
b. Is X continuous or discrete?
c. X ~ ________
d. On average, how long would you expect one car battery to last?
e. On average, how long would you expect nine car batteries to last, if they are used one after another?
f. Find the probability that a car battery lasts more than 36 months.
g. Seventy percent of the batteries last at least how long?

Solution:

a. X = the useful life of a particular car battery, measured in months.

b. X is continuous.
c. $X \sim Exp(0.025)$
d. 40 months
e. 360 months
f. 0.4066
g. 14.27

Exercise:

Problem:

The percent of persons (ages five and older) in each state who speak a language at home other than English is approximately exponentially distributed with a mean of 9.848. Suppose we randomly pick a state.

a. Define the random variable. X = ______________________________________.
b. Is X continuous or discrete?
c. $X \sim$ _________
d. μ = _________
e. σ = _________
f. Draw a graph of the probability distribution. Label the axes.
g. Find the probability that the percent is less than 12.
h. Find the probability that the percent is between eight and 14.
i. The percent of all individuals living in the United States who speak a language at home other than English is 13.8.
    i. Why is this number different from 9.848%?
    ii. What would make this number higher than 9.848%?

Exercise:

Problem:

The time (in years) **after** reaching age 60 that it takes an individual to retire is approximately exponentially distributed with a mean of about five years. Suppose we randomly pick one retired individual. We are interested in the time after age 60 to retirement.

a. Define the random variable. X = ______________________________________.
b. Is X continuous or discrete?
c. X ~ = _________
d. μ = _________
e. σ = _________
f. Draw a graph of the probability distribution. Label the axes.
g. Find the probability that the person retired after age 70.
h. Do more people retire before age 65 or after age 65?
i. In a room of 1,000 people over age 80, how many do you expect will NOT have retired yet?

Solution:

a. X = the time (in years) after reaching age 60 that it takes an individual to retire
b. X is continuous.
c. $X \sim Exp\left(\frac{1}{5}\right)$
d. five
e. five
f. Check student's solution.
g. 0.1353
h. before
i. 18.3

Exercise:

Problem:

The cost of all maintenance for a car during its first year is approximately exponentially distributed with a mean of $150.

a. Define the random variable. X = ______________________________________.
b. X ~ = _________
c. μ = _________
d. σ = _________
e. Draw a graph of the probability distribution. Label the axes.

f. Find the probability that a car required over $300 for maintenance during its first year.

Use the following information to answer the next three exercises. The average lifetime of a certain new cell phone is three years. The manufacturer will replace any cell phone failing within two years of the date of purchase. The lifetime of these cell phones is known to follow an exponential distribution.

Exercise:

Problem: The decay rate is:

a. 0.3333
b. 0.5000
c. 2
d. 3

Solution:

a

Exercise:

Problem:

What is the probability that a phone will fail within two years of the date of purchase?

a. 0.8647
b. 0.4866
c. 0.2212
d. 0.9997

Exercise:

Problem: What is the median lifetime of these phones (in years)?

a. 0.1941

b. 1.3863
c. 2.0794
d. 5.5452

Solution:

c

Exercise:

Problem: Let X ~ *Exp*(0.1).

a. decay rate = ________
b. μ = ________
c. Graph the probability distribution function.
d. On the graph, shade the area corresponding to $P\left(x < 6\right)$ and find the probability.
e. Sketch a new graph, shade the area corresponding to $P\left(3 < x < 6\right)$ and find the probability.
f. Sketch a new graph, shade the area corresponding to $P\left(x < 7\right)$ and find the probability.
g. Sketch a new graph, shade the area corresponding to the 40th percentile and find the value.
h. Find the average value of X.

Exercise:

Problem:

Suppose that the longevity of a light bulb is exponential with a mean lifetime of eight years.

a. Find the probability that a light bulb lasts less than one year.
b. Find the probability that a light bulb lasts between six and ten years.
c. Seventy percent of all light bulbs last at least how long?
d. A company decides to offer a warranty to give refunds to light bulbs whose lifetime is among the lowest two percent of all bulbs. To the nearest month, what should be the cutoff lifetime for the warranty to take place?

e. If a light bulb has lasted seven years, what is the probability that it fails within the 8th year.

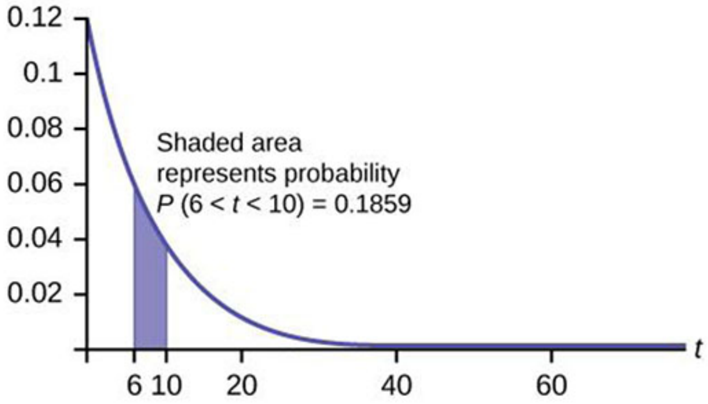
Solution:

Let T = the life time of a light bulb.

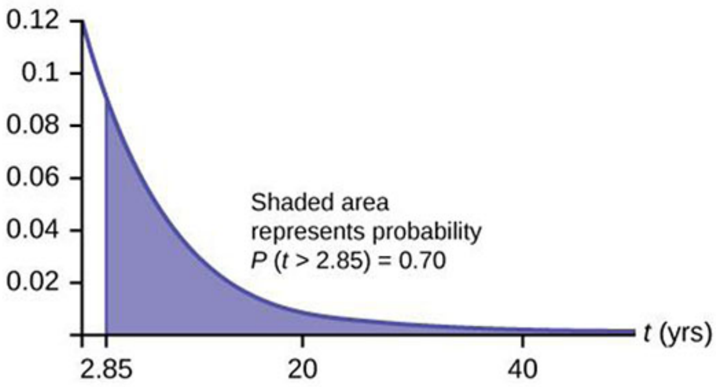
The decay parameter is m = 1/8, and $T\sim$ Exp(1/8). The cumulative distribution function is $P\left(T<t\right)=1-e^{-\frac{1}{8}t}$

a. Therefore, $P\left(t<1\right)=1-e^{-\frac{1}{8}}=0.1175.$
b. We want to find $P\left(6<t<10\right)$. To do this,

$$P\left(6<t<10\right)=P\left(t<10\right)-P\left(t<6\right)$$
$$=\left(1-e^{-\frac{1}{8}(10)}\right)-\left(1-e^{-\frac{1}{8}(6)}\right)=0.7135-0.5276=0.1859$$



c. We want to find $0.70=P(T>t)=1–\left(1–e^{-\frac{t}{8}}\right)=e^{-\frac{t}{8}}$.
Solving for t, $e^{-\frac{t}{8}}$ = 0.70, so $-\frac{t}{8}=\ln\left(0.70\right)$, and
$t=–8\ln(0.70)\approx2.85$ years.



d. We want to find $0.02 = P\left(T < t\right) = 1 - e^{-\frac{1}{8}t}$.
Solving for t, $e^{-\frac{t}{8}} = 0.98$, so $-\frac{t}{8} = \ln\left(0.98\right)$, and
$t = -8\ln(0.98) \approx 0.1616$ years, or roughly two months.

The warranty should cover light bulbs that last less than 2 months.

e. We must find $P\left(t < 8 | t > 7\right)$.

Notice that by the rule of complement events,
$P\left(t < 8 | t > 7\right) = 1 - P\left(t > 8 | t > 7\right)$.

By the memoryless property ($P\left(t > 8 | t > 7\right) = P\left(t > 1\right)$ and
$P\left(t > 1\right) = 1 - \left(1 - e^{-\frac{1}{8}}\right) = e^{-\frac{1}{8}} \approx 0.8825$

Therefore, $P\left(t < 8 | t > 7\right) = 1 - 0.8825 = 0.1175$.

Exercise:

Problem:

At a 911 call center, calls come in at an average rate of one call every two minutes. Assume that the time that elapses from one call to the next has the exponential distribution.

a. On average, how much time occurs between five consecutive calls?
b. Find the probability that after a call is received, it takes more than three minutes for the next call to occur.
c. Ninety-percent of all calls occur within how many minutes of the previous call?

d. Suppose that two minutes have elapsed since the last call. Find the probability that the next call will occur within the next minute.
e. Find the probability that less than 20 calls occur within an hour.

Exercise:

Problem:

At an urgent care facility, patients arrive at an average rate of one patient every seven minutes. Assume that the duration between arrivals is exponentially distributed.

a. Find the probability that the time between two successive visits to the urgent care facility is less than 2 minutes.
b. Find the probability that the time between two successive visits to the urgent care facility is more than 15 minutes.
c. If 10 minutes have passed since the last arrival, what is the probability that the next person will arrive within the next five minutes?

Solution:

Let T = duration (in minutes) between successive visits. Since patients arrive at a rate of one patient every seven minutes, μ = 7 and the decay constant is $m = \frac{1}{7}$. The cdf is $P\left(T < t\right) = 1 - e^{-\frac{1}{7}t}$

a. $P\left(t < 2\right) = 1 - e^{-\frac{1}{7}(2)} = 0.2485.$
b.
$$P\left(t > 15\right) = 1 - P\left(t < 15\right) = 1 - \left(1 - e^{-\frac{15}{7}}\right) \approx e^{-\frac{15}{7}} \approx 0.1173$$
.
c.
$$P\left(t > 15|t > 10\right) = P\left(t > 5\right) = 1 - \left(1 - e^{-\frac{5}{7}}\right) = e^{-\frac{5}{7}} = 0.4895$$
.

Glossary

decay parameter

The decay parameter describes the rate at which probabilities decay to zero for increasing values of x. It is the value m in the probability density function $f\left(x\right)=me^{-mx}$ of an exponential random variable. It is also equal to $m=\frac{1}{\mu}$, where μ is the mean of the random variable.

memoryless property

For an exponential random variable X, the memoryless property is the statement that knowledge of what has occurred in the past has no effect on future probabilities. This means that the probability that x exceeds $t+r$, given that it has exceeded t, is the same as the probability that x would exceed r if we had no knowledge about it. In symbols we say that $P\left(x>t+r|x>t\right)=P\left(x>r\right)$.

The Normal Distribution: Introduction

class="introduction"

If you ask enough people about their shoe size, you will find that your graphed data is shaped like a bell curve and can be described as normally distributed. (credit: Ömer Ünlü)

Note:

Chapter Objectives

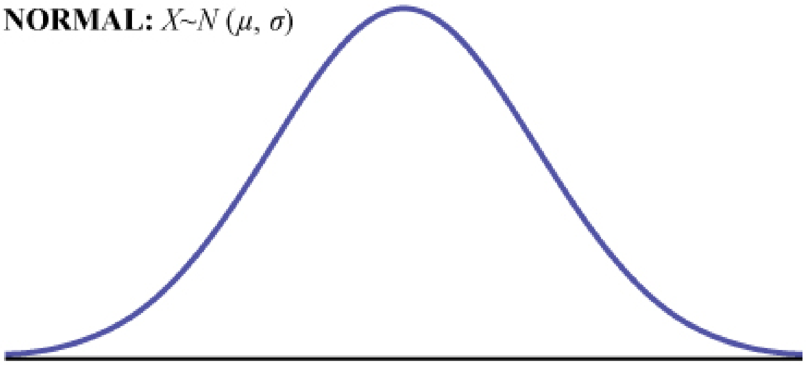
By the end of this chapter, the student should be able to:

- Recognize the normal probability distribution and apply it appropriately.
- Recognize the standard normal probability distribution and apply it appropriately.
- Compare normal probabilities by converting to the standard normal distribution.

The normal, a continuous distribution, is the most important of all the distributions. It is widely used and even more widely abused. Its graph is bell-shaped. You see the bell curve in almost all disciplines. Some of these include psychology, business, economics, the sciences, nursing, and, of course, mathematics. Some of your instructors may use the normal distribution to help determine your grade. Most IQ scores are normally distributed. Often real-estate prices fit a normal distribution. The normal distribution is extremely important, but it cannot be applied to everything in the real world.

In this chapter, you will study the normal distribution, the standard normal distribution, and applications associated with them.

The normal distribution has two parameters (two numerical descriptive measures), the mean (μ) and the standard deviation (σ). If X is a quantity to be measured that has a normal distribution with mean (μ) and standard deviation (σ), we designate this by writing



The probability density function is a rather complicated function. **Do not memorize it**. It is not necessary.

$$f(x) = \frac{1}{\sigma \cdot \sqrt{2 \cdot \pi}} \cdot \mathrm{e}^{-\frac{1}{2} \cdot \left(\frac{x-\mu}{\sigma}\right)^2}$$

The cumulative distribution function is $P(X < x)$. It is calculated either by a calculator or a computer, or it is looked up in a table. Technology has made the tables virtually obsolete. For that reason, as well as the fact that there are various table formats, we are not including table instructions.

The curve is symmetrical about a vertical line drawn through the mean, μ. In theory, the mean is the same as the median, because the graph is symmetric about μ. As the notation indicates, the normal distribution depends only on the mean and the standard deviation. Since the area under the curve must equal one, a change in the standard deviation, σ, causes a change in the shape of the curve; the curve becomes fatter or skinnier depending on σ. A change in μ causes the graph to shift to the left or right. This means there are an infinite number of normal probability distributions. One of special interest is called the **standard normal distribution**.

Note:
Collaborative Classroom Activity
Your instructor will record the heights of both men and women in your class, separately. Draw histograms of your data. Then draw a smooth curve through each histogram. Is each curve somewhat bell-shaped? Do you think that if you had recorded 200 data values for men and 200 for women that the curves would look bell-shaped? Calculate the mean for each data set. Write the means on the x-axis of the appropriate graph below the peak. Shade the approximate area that represents the probability that one randomly chosen male is taller than 72 inches. Shade the approximate area that represents the probability that one randomly chosen female is shorter than 60 inches. If the total area under each curve is one, does either probability appear to be more than 0.5?

Formula Review

$X \sim N(\mu, \sigma)$

μ = the mean σ = the standard deviation

Glossary

Normal Distribution
a continuous random variable (RV) with pdf $f(x)$ =

$$\frac{1}{\sigma\sqrt{2\pi}}\,\mathrm{e}^{{\frac{-(x-m)}{2\sigma^{2}}}^{2}}$$

, where μ is the mean of the distribution and σ is the standard deviation; notation: X ~ $N(\mu, \sigma)$. If μ = 0 and σ = 1, the RV is called the **standard normal distribution**.

The Standard Normal Distribution

The **standard normal distribution** is a normal distribution of **standardized values called z-scores**. **A z-score is measured in units of the standard deviation.** For example, if the mean of a normal distribution is five and the standard deviation is two, the value 11 is three standard deviations above (or to the right of) the mean. The calculation is as follows:

$x = \mu + (z)(\sigma) = 5 + (3)(2) = 11$

The z-score is three.

The mean for the standard normal distribution is zero, and the standard deviation is one. The transformation $z = \frac{x-\mu}{\sigma}$ produces the distribution $Z \sim N(0,1)$. The value x comes from a normal distribution with mean μ and standard deviation σ.

Z-Scores

If X is a normally distributed random variable and $X \sim N(\mu, \sigma)$, then the z-score is:

Equation:

$$z = \frac{x - \mu}{\sigma}$$

The z-score tells you how many standard deviations the value x is above (to the right of) or below (to the left of) the mean, μ. Values of x that are larger than the mean have positive z-scores, and values of x that are smaller than the mean have negative z-scores. If x equals the mean, then x has a z-score of zero.

Example:
Suppose $X \sim N(5, 6)$. This says that x is a normally distributed random variable with mean μ = 5 and standard deviation σ = 6. Suppose x = 17. Then:

Equation:

$$z = \frac{x - \mu}{\sigma} = \frac{17 - 5}{6} = 2$$

This means that x = 17 is **two standard deviations** (2σ) above or to the right of the mean μ = 5. The standard deviation is σ = 6.

Notice that: 5 + (2)(6) = 17 (The pattern is $\mu + z\sigma = x$)

Now suppose x = 1. Then: $z = \frac{x - \mu}{\sigma} = \frac{1 - 5}{6}$ = –0.67 (rounded to two decimal places)

This means that x = 1 is 0.67 standard deviations (–0.67σ) below or to the left of the mean μ = 5. Notice that: 5 + (–0.67)(6) is approximately equal to one (This has the pattern $\mu + (-0.67)\sigma = 1$)

Summarizing, when z is positive, x is above or to the right of μ and when z is negative, x is to the left of or below μ. Or, when z is positive, x is greater than μ, and when z is negative x is less than μ.

Note:
Try It

Exercise:

Problem: What is the z-score of x, when x = 1 and $X \sim N(12, 3)$?

Solution:

$$z = \frac{1 - 12}{3} \approx -3.67$$

Example:

Some doctors believe that a person can lose five pounds, on the average, in a month by reducing his or her fat intake and by exercising consistently. Suppose weight loss has a normal distribution. Let X = the amount of weight lost (in pounds) by a person in a month. Use a standard deviation of two pounds. $X \sim N(5, 2)$. Fill in the blanks.

Exercise:

Problem:

a. Suppose a person **lost** ten pounds in a month. The z-score when x = 10 pounds is z = 2.5 (verify). This z-score tells you that x = 10 is ________ standard deviations to the ________ (right or left) of the mean _____ (What is the mean?).

Solution:

a. This z-score tells you that x = 10 is **2.5** standard deviations to the **right** of the mean **five**.

Exercise:

Problem:

b. Suppose a person **gained** three pounds (a negative weight loss). Then z = __________. This z-score tells you that x = –3 is ________ standard deviations to the __________ (right or left) of the mean.

Solution:

b. z = **–4**. This z-score tells you that x = –3 is **four** standard deviations to the **left** of the mean.

Suppose the random variables X and Y have the following normal distributions: $X \sim N(5, 6)$ and $Y \sim N(2, 1)$. If x = 17, then z = 2. (This was previously shown.) If y = 4, then what is z?

$z = \frac{y-\mu}{\sigma} = \frac{4-2}{1} = 2$ where $\mu = 2$ and $\sigma = 1$.

The z-score for $y = 4$ is $z = 2$. This means that four is $z = 2$ standard deviations to the right of the mean. Therefore, $x = 17$ and $y = 4$ are both two (of **their own**) standard deviations to the right of **their** respective means.

The z-score allows us to compare data that are scaled differently. To understand the concept, suppose $X \sim N(5, 6)$ represents weight gains for one group of people who are trying to gain weight in a six week period and $Y \sim N(2, 1)$ measures the same weight gain for a second group of people. A negative weight gain would be a weight loss. Since $x = 17$ and $y = 4$ are each two standard deviations to the right of their means, they represent the same, standardized weight gain **relative to their means**.

Note:
Try It
Exercise:

Problem: Fill in the blanks.

Jerome averages 16 points a game with a standard deviation of four points. $X \sim N(16, 4)$. Suppose Jerome scores ten points in a game. The z–score when $z = 10$ is –1.5. This score tells you that $x = 10$ is ______ standard deviations to the ______(right or left) of the mean______(What is the mean?).

Solution:

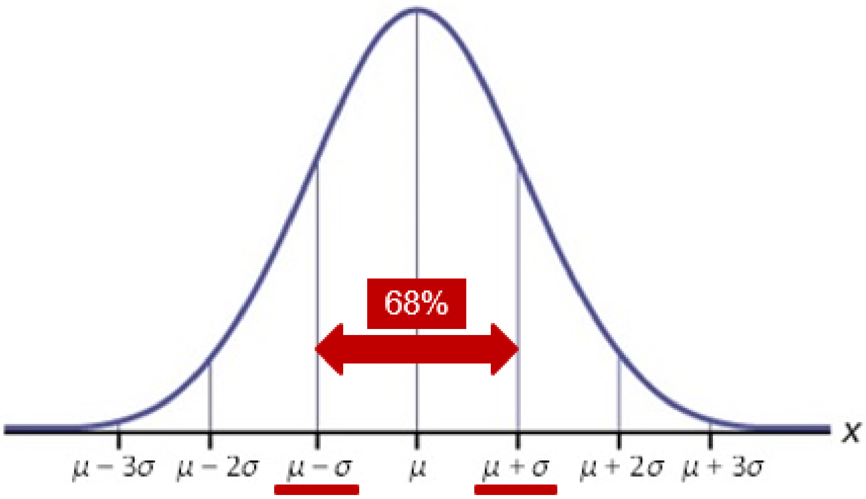
1.5, left, 16

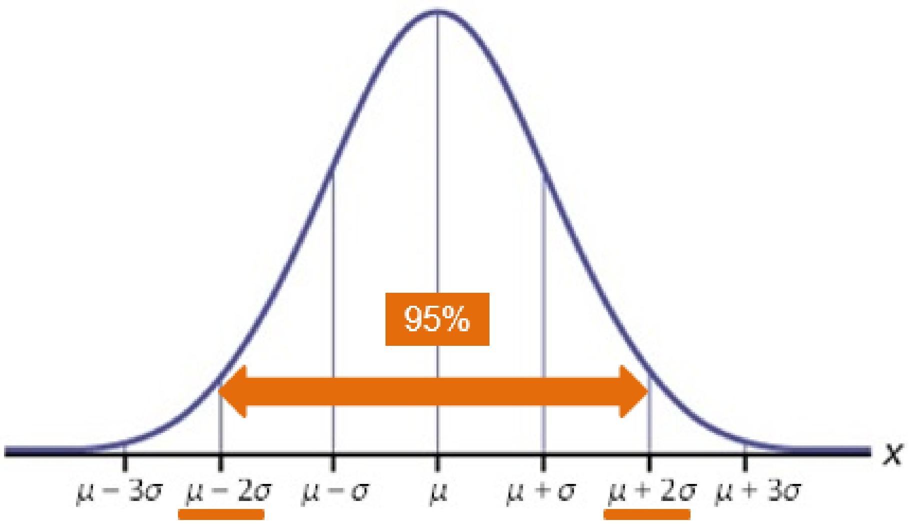
The Empirical Rule

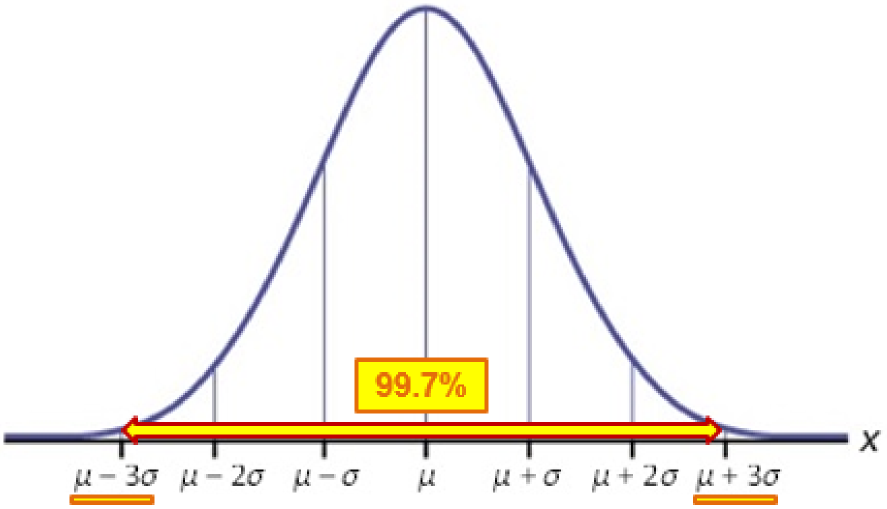
If X is a random variable and has a normal distribution with mean μ and standard deviation σ, then the **Empirical Rule** says the following:

- About 68% of the x values lie between –1σ and +1σ of the mean μ (within one standard deviation of the mean).
- About 95% of the x values lie between –2σ and +2σ of the mean μ (within two standard deviations of the mean).
- About 99.7% of the x values lie between –3σ and +3σ of the mean μ (within three standard deviations of the mean). Notice that almost all the x values lie within three standard deviations of the mean.
- The z-scores for +1σ and –1σ are +1 and –1, respectively.
- The z-scores for +2σ and –2σ are +2 and –2, respectively.
- The z-scores for +3σ and –3σ are +3 and –3 respectively.

The Empirical Rule is also known as the 68-95-99.7 rule and is illustrated in the following graphs.







Example:

The mean height of 15 to 18-year-old males from Chile from 2009 to 2010 was 170 cm with a standard deviation of 6.28 cm. Male heights are known to follow a normal distribution. Let X = the height of a 15 to 18-year-old male from Chile in 2009 to 2010. Then $X \sim N(170, 6.28)$.

Exercise:

Problem:

a. Suppose a 15 to 18-year-old male from Chile was 168 cm tall from 2009 to 2010. The z-score when x = 168 cm is z = _______. This z-score tells you that x = 168 is ________ standard deviations to the ________ (right or left) of the mean _____ (What is the mean?).

Solution:

a. –0.32, 0.32, left, 170

Exercise:

Problem:

b. Suppose that the height of a 15 to 18-year-old male from Chile from 2009 to 2010 has a z-score of z = 1.27. What is the male's height? _______ The z-score (z = 1.27) tells you that the male's height is ________ standard deviations to the __________ (right or left) of the mean.

Solution:

b. 177.98, 1.27, right

Note:
Try It
Exercise:

Problem:

Use the information in [link] to answer the following questions.

a. Suppose a 15 to 18-year-old male from Chile was 176 cm tall from 2009 to 2010. The z-score when x = 176 cm is z = _______. This z-score tells you that x = 176 cm is ________ standard deviations to the ________ (right or left) of the mean _____ (What is the mean?).

b. Suppose that the height of a 15 to 18-year-old male from Chile from 2009 to 2010 has a z-score of z = –2. What is the male's height? ______ The z-score (z = –2) tells you that the male's height is ________ standard deviations to the __________ (right or left) of the mean.

Solution:

Solve the equation $z = \frac{x-\mu}{\sigma}$ for x. $x = \mu + (z)(\sigma)$

a. $z = \frac{176-170}{6.28} \approx 0.96$, This z-score tells you that x = 176 cm is 0.96 standard deviations to the right of the mean 170 cm.
b. x = 157.44 cm, The z-score (z = –2) tells you that the male's height is two standard deviations to the left of the mean.

Example:
Exercise:

Problem:

From 1984 to 1985, the mean height of 15 to 18-year-old males from Chile was 172.36 cm, and the standard deviation was 6.34 cm. Let Y = the height of 15 to 18-year-old males from 1984 to 1985. Then $Y \sim N(172.36, 6.34)$.

The mean height of 15 to 18-year-old males from Chile from 2009 to 2010 was 170 cm with a standard deviation of 6.28 cm. Male heights are known to follow a normal distribution. Let X = the height of a 15 to 18-year-old male from Chile in 2009 to 2010. Then $X \sim N(170, 6.28)$.

Find the z-scores for x = 160.58 cm and y = 162.85 cm. Interpret each z-score. What can you say about x = 160.58 cm and y = 162.85 cm?

Solution:

The z-score for x = 160.58 is z = –1.5.
The z-score for y = 162.85 is z = –1.5.
Both x = 160.58 and y = 162.85 deviate the same number of standard deviations from their respective means and in the same direction.

Note:
Try It
Exercise:

Problem:

In 2015, the distribution of scores in the mathematics section of the SAT exam had a mean μ = 511 and a standard deviation σ = 210. Let X = a SAT exam mathematics section score in 2015. Then X ~ $N(511, 210)$.

In the same year, the distribution of scores in the mathematics section of the ACT exam had a mean μ = 20.8 and a standard deviation σ = 5.4. Let Y = an ACT exam mathematics section score in 2015. Then Y ~ $N(20.8, 5.4)$.

Find the z-scores for SAT score x = 325 and ACT score y = 19. Interpret each z-score. What can you say about x = 325 and y = 19?

Solution:

The z-score for x = 325 is z = –0.89. This means that a score of 325 is about 0.89 standard deviations below the mean.

The z-score for y = 19 is z = –0.33. This means that a score of 19 is about 0.33 standard deviations below the mean.

The student who scored a 19 on the mathematics section of the ACT exam scored closer to the mean than the student who scored a 325 on the mathematics section of the SAT exam and the ACT score is the better score since it has a higher z-score. Since both exams are measuring college readiness, it can be concluded that the student who scored a 19 on the mathematics section of the ACT test appears to be more prepared for college mathematics than the student who scored a 325 on the mathematics section of the SAT exam.

Example:

Suppose X has a normal distribution with mean 50 and standard deviation 6. Since this is a normal distribution, we know by the **Empirical Rule** the following must be true:

- About 68% of the x values lie between –1σ = (–1)(6) = –6 and 1σ = (1)(6) = 6 of the mean 50. The values 50 – 6 = 44 and 50 + 6 = 56 are within one standard deviation of the mean 50. The z-scores are –1 and +1 for 44 and 56, respectively.
- About 95% of the x values lie between –2σ = (–2)(6) = –12 and 2σ = (2)(6) = 12. The values 50 – 12 = 38 and 50 + 12 = 62 are within two standard deviations of the mean 50. The z-scores are –2 and +2 for 38 and 62, respectively.
- About 99.7% of the x values lie between –3σ = (–3)(6) = –18 and 3σ = (3)(6) = 18 of the mean 50. The values 50 – 18 = 32 and 50 + 18 = 68 are within three standard deviations of the mean 50. The z-scores are –3 and +3 for 32 and 68, respectively.

Note:

Try It

Exercise:

Problem:

Suppose X has a normal distribution with mean 25 and standard deviation five. Between what values of x do 68% of the values lie?

Solution:

between 20 and 30.

Example:
Exercise:

Problem:

From 1984 to 1985, the mean height of 15 to 18-year-old males from Chile was 172.36 cm, and the standard deviation was 6.34 cm. Let Y = the height of 15 to 18-year-old males in 1984 to 1985. Then Y ~ $N(172.36, 6.34)$.

a. About 68% of the y values lie between what two values? These values are __________________. The z-scores are __________________, respectively.
b. About 95% of the y values lie between what two values? These values are __________________. The z-scores are __________________ respectively.
c. About 99.7% of the y values lie between what two values? These values are __________________. The z-scores are __________________, respectively.

Solution:

a. About 68% of the values lie between 166.02 and 178.7. The z-scores are –1 and 1.

b. About 95% of the values lie between 159.68 and 185.04. The z-scores are –2 and 2.
c. About 99.7% of the values lie between 153.34 and 191.38. The z-scores are –3 and 3.

Note:
Try It
Exercise:

Problem:

The scores on a college entrance exam have an approximate normal distribution with mean, μ = 52 points and a standard deviation, σ = 11 points.

a. About 68% of the x values lie between what two values? These values are _________________. The z-scores are _________________, respectively.
b. About 95% of the x values lie between what two values? These values are _________________. The z-scores are _________________, respectively.
c. About 99.7% of the x values lie between what two values? These values are _________________. The z-scores are _________________, respectively.

Solution:

a. About 68% of the values lie between the values 41 and 63. The z-scores are –1 and 1, respectively.
b. About 95% of the values lie between the values 30 and 74. The z-scores are –2 and 2, respectively.
c. About 99.7% of the values lie between the values 19 and 85. The z-scores are –3 and 3, respectively.

References

“Blood Pressure of Males and Females.” StatCruch, 2013. Available online at http://www.statcrunch.com/5.0/viewreport.php?reportid=11960 (accessed May 14, 2013).

“The Use of Epidemiological Tools in Conflict-affected populations: Open-access educational resources for policy-makers: Calculation of z-scores.” London School of Hygiene and Tropical Medicine, 2009. Available online at http://conflict.lshtm.ac.uk/page_125.htm (accessed May 14, 2013).

“2015 College-Bound Seniors Total Group Profile Report.” CollegeBoard, 2015. Available online at https://secure-media.collegeboard.org/digitalServices/pdf/sat/total-group-2015.pdf (accessed June 29, 2017).

“Digest of Education Statistics: Number and percentage of graduates taking the ACT test; average scores and standard deviations, by sex and race/ethnicity; and percentage of test takers with selected composite scores and planned fields of postsecondary study: Selected years, 1995 through 2015” National Center for Education Statistics. Available online at https://nces.ed.gov/programs/digest/d15/tables/dt15_226.50.asp (accessed June 29, 2017).

Data from the *San Jose Mercury News*.

Data from *The World Almanac and Book of Facts*.

“List of stadiums by capacity.” Wikipedia. Available online at https://en.wikipedia.org/wiki/List_of_stadiums_by_capacity (accessed May 14, 2013).

Data from the National Basketball Association. Available online at www.nba.com (accessed May 14, 2013).

Chapter Review

A z-score is a standardized value. Its distribution is the standard normal, Z ~ $N(0, 1)$. The mean of the z-scores is zero and the standard deviation is one. If z is the z-score for a value x from the normal distribution $N(\mu, \sigma)$ then z tells you how many standard deviations x is above (greater than) or below (less than) μ.

Formula Review

Z ~ $N(0, 1)$

z = a standardized value (z-score)

mean = 0; standard deviation = 1

z-score: $z = \frac{x - \mu}{\sigma}$

Z = the random variable for z-scores

Z ~ $N(0, 1)$

Exercise:

Problem:

A bottle of water contains 12.05 fluid ounces with a standard deviation of 0.01 ounces. Define the random variable X in words. X = ____________.

Solution:

ounces of water in a bottle

Exercise:

Problem:

A normal distribution has a mean of 61 and a standard deviation of 15. What is the median?

Exercise:

Problem: $X \sim N(1, 2)$

σ = _______

Solution:

2

Exercise:

Problem:

A company manufactures rubber balls. The mean diameter of a ball is 12 cm with a standard deviation of 0.2 cm. Define the random variable X in words. X = ______________.

Exercise:

Problem: $X \sim N(-4, 1)$

What is the median?

Solution:

–4

Exercise:

Problem: $X \sim N(3, 5)$

σ = _______

Exercise:

Problem: $X \sim N(-2, 1)$

μ = _______

Solution:

–2

Exercise:

Problem: What does a z-score measure?

Exercise:

Problem:

What does standardizing a normal distribution do to the mean?

Solution:

The mean becomes zero.

Exercise:

Problem:

Is $X \sim N(0, 1)$ a standardized normal distribution? Why or why not?

Exercise:

Problem:

What is the z-score of $x = 12$, if it is two standard deviations to the right of the mean?

Solution:

$z = 2$

Exercise:

Problem:

What is the z-score of $x = 9$, if it is 1.5 standard deviations to the left of the mean?

Exercise:

Problem:

What is the z-score of x = –2, if it is 2.78 standard deviations to the left of the mean?

Solution:

z = -2.78

Exercise:

Problem:

What is the z-score of x = 7, if it is 0.133 standard deviations to the right of the mean?

Exercise:

Problem:

Suppose X ~ $N(2, 6)$. What value of x has a z-score of three?

Solution:

x = 20

Exercise:

Problem:

Suppose X ~ $N(8, 1)$. What value of x has a z-score of –2.25?

Exercise:

Problem:

Suppose X ~ $N(9, 5)$. What value of x has a z-score of –0.5?

Solution:

x = 6.5

Exercise:

Problem:

Suppose $X \sim N(2, 3)$. What value of x has a z-score of –0.67?

Exercise:

Problem:

Suppose $X \sim N(4, 2)$. What value of x is 1.5 standard deviations to the left of the mean?

Solution:

$x = 1$

Exercise:

Problem:

Suppose $X \sim N(4, 2)$. What value of x is two standard deviations to the right of the mean?

Exercise:

Problem:

Suppose $X \sim N(8, 9)$. What value of x is 0.67 standard deviations to the left of the mean?

Solution:

$x = 1.97$

Exercise:

Problem: Suppose $X \sim N(-1, 2)$. What is the z-score of $x = 2$?

Exercise:

Problem: Suppose $X \sim N(12, 6)$. What is the z-score of $x = 2$?

Solution:

z = –1.67

Exercise:

Problem: Suppose $X \sim N(9, 3)$. What is the z-score of x = 9?

Exercise:

Problem:

Suppose a normal distribution has a mean of six and a standard deviation of 1.5. What is the z-score of x = 5.5?

Solution:

z ≈ –0.33

Exercise:

Problem:

In a normal distribution, x = 5 and z = –1.25. This tells you that x = 5 is _____ standard deviations to the _____ (right or left) of the mean.

Exercise:

Problem:

In a normal distribution, x = 3 and z = 0.67. This tells you that x = 3 is _____ standard deviations to the _____ (right or left) of the mean.

Solution:

0.67, right

Exercise:

Problem:

In a normal distribution, x = –2 and z = 6. This tells you that x = –2 is ____ standard deviations to the ____ (right or left) of the mean.

Exercise:

Problem:

In a normal distribution, x = –5 and z = –3.14. This tells you that x = –5 is ____ standard deviations to the ____ (right or left) of the mean.

Solution:

3.14, left

Exercise:

Problem:

In a normal distribution, x = 6 and z = –1.7. This tells you that z = 6 is ____ standard deviations to the ____ (right or left) of the mean.

Exercise:

Problem:

About what percent of x values from a normal distribution lie within one standard deviation (left and right) of the mean of that distribution?

Solution:

about 68%

Exercise:

Problem:

About what percent of the x values from a normal distribution lie within two standard deviations (left and right) of the mean of that distribution?

Exercise:

Problem:

About what percent of x values lie between the second and third standard deviations (both sides)?

Solution:

about 4.7%

Exercise:

Problem:

Suppose $X \sim N(15, 3)$. Between what x values does about 68% of the data lie? The range of x values is centered at the mean of the distribution (i.e., 15).

Exercise:

Problem:

Suppose $X \sim N(-3, 1)$. Between what x values does about 95% of the data lie? The range of x values is centered at the mean of the distribution (i.e., –3).

Solution:

between –5 and –1

Exercise:

Problem:

Suppose $X \sim N(-3, 1)$. Between what x values does about 34% of the data lie?

Exercise:

Problem:

About what percent of x values lie between the mean and three standard deviations (one side)?

Solution:

about 50%

Exercise:

Problem:

About what percent of x values lie between the mean and one standard deviation (one side)?

Exercise:

Problem:

About what percent of x values lie between the first and second standard deviations from the mean (both sides)?

Solution:

about 27%

Exercise:

Problem:

About what percent of x values lie between the first and third standard deviations (both sides)?

Use the following information to answer the next two exercises: The life of Sunshine CD players is normally distributed with mean of 4.1 years and a standard deviation of 1.3 years. A CD player is guaranteed for three years. We are interested in the length of time a CD player lasts.

Exercise:

Problem:

Define the random variable X in words. X = _______________.

Solution:

The lifetime of a Sunshine CD player measured in years.

Exercise:

Problem: X ~ _____(_____,_____)

Homework

Use the following information to answer the next two exercises: The patient recovery time from a particular surgical procedure is normally distributed with a mean of 5.3 days and a standard deviation of 2.1 days.

Exercise:

Problem: What is the median recovery time?

a. 2.7
b. 5.3
c. 7.4
d. 2.1

Exercise:

Problem:

What is the z-score for a patient who takes ten days to recover?

a. 1.5
b. 0.2
c. 2.2
d. 7.3

Solution:

c

Exercise:

Problem:

The length of time to find it takes to find a parking space at 9 A.M. follows a normal distribution with a mean of five minutes and a standard deviation of two minutes. If the mean is significantly greater than the standard deviation, which of the following statements is true?

I. The data cannot follow the uniform distribution.
II. The data cannot follow the exponential distribution..
III. The data cannot follow the normal distribution.

a. I only
b. II only
c. III only
d. I, II, and III

Exercise:

Problem:

The heights of the 430 National Basketball Association players were listed on team rosters at the start of the 2005–2006 season. The heights of basketball players have an approximate normal distribution with mean, $\mu = 79$ inches and a standard deviation, $\sigma = 3.89$ inches. For each of the following heights, calculate the z-score and interpret it using complete sentences.

a. 77 inches
b. 85 inches
c. If an NBA player reported his height had a z-score of 3.5, would you believe him? Explain your answer.

Solution:

a. Use the z-score formula. $z = –0.5141$. The height of 77 inches is 0.5141 standard deviations below the mean. An NBA player whose height is 77 inches is shorter than average.

b. Use the z-score formula. z = 1.5424. The height 85 inches is 1.5424 standard deviations above the mean. An NBA player whose height is 85 inches is taller than average.
c. Height = 79 + 3.5(3.89) = 90.67 inches, which is over 7.7 feet tall. There are very few NBA players this tall so the answer is no, not likely.

Exercise:

Problem:

The systolic blood pressure (given in millimeters) of males has an approximately normal distribution with mean μ = 125 and standard deviation σ = 14. Systolic blood pressure for males follows a normal distribution.

a. Calculate the z-scores for the male systolic blood pressures 100 and 150 millimeters.
b. If a male friend of yours said he thought his systolic blood pressure was 2.5 standard deviations below the mean, but that he believed his blood pressure was between 100 and 150 millimeters, what would you say to him?

Exercise:

Problem:

Kyle's doctor told him that the z-score for his systolic blood pressure is 1.75. Which of the following is the best interpretation of this standardized score? The systolic blood pressure (given in millimeters) of males has an approximately normal distribution with mean μ = 125 and standard deviation σ = 14. If X = a systolic blood pressure score, then X ~ $N(125, 14)$.

a. Which answer(s) **is/are** correct?

    i. Kyle's systolic blood pressure is 175.

   ii. Kyle’s systolic blood pressure is 1.75 times the average blood pressure of men his age.
   iii. Kyle’s systolic blood pressure is 1.75 above the average systolic blood pressure of men his age.
   iv. Kyles’s systolic blood pressure is 1.75 standard deviations above the average systolic blood pressure for men.

b. Calculate Kyle’s blood pressure.

Solution:

a. iv
b. Kyle’s blood pressure is equal to 125 + (1.75)(14) = 149.5.

Exercise:

Problem:

Height and weight are two measurements used to track a child’s development. The World Health Organization measures child development by comparing the weights of children who are the same height and the same gender. In 2009, weights for all 80 cm girls in the reference population had a mean μ = 10.2 kg and standard deviation σ = 0.8 kg. Weights are normally distributed. $X \sim N(10.2, 0.8)$. Calculate the z-scores that correspond to the following weights and interpret them.

a. 11 kg
b. 7.9 kg
c. 12.2 kg

Exercise:

Problem:

In 2005, 1,475,623 students heading to college took the SAT. The distribution of scores in the math section of the SAT follows a normal distribution with mean μ = 520 and standard deviation σ = 115.

a. Calculate the z-score for a SAT score of 720. Interpret it using a complete sentence.
b. What math SAT score is 1.5 standard deviations above the mean? What can you say about this SAT score?
c. For 2012, the SAT math test had a mean of 514 and standard deviation 117. The ACT math test is an alternate to the SAT and is approximately normally distributed with mean 21 and standard deviation 5.3. If one person took the SAT math test and scored 700 and a second person took the ACT math test and scored 30, who did better with respect to the test they took?

Solution:

Let X = a SAT math score and Y = an ACT math score.

a. x = 720; $z = \frac{720 - 520}{115}$ = 1.74. The exam score of 720 is 1.74 standard deviations above the mean of 520.
b. z = 1.5. The math SAT score is 520 + 1.5(115) ≈ 692.5. The exam score of 692.5 is 1.5 standard deviations above the mean of 520.
c. $z = \frac{x - \mu}{\sigma} = \frac{700 - 514}{117}$ ≈ 1.59, the z-score for the SAT.

   $z = \frac{y - \mu}{\sigma} = \frac{30 - 21}{5.3}$ ≈ 1.70, the z-score for the ACT.

   With respect to the test they took, the person who took the ACT did better (has the higher z-score).

Glossary

Standard Normal Distribution

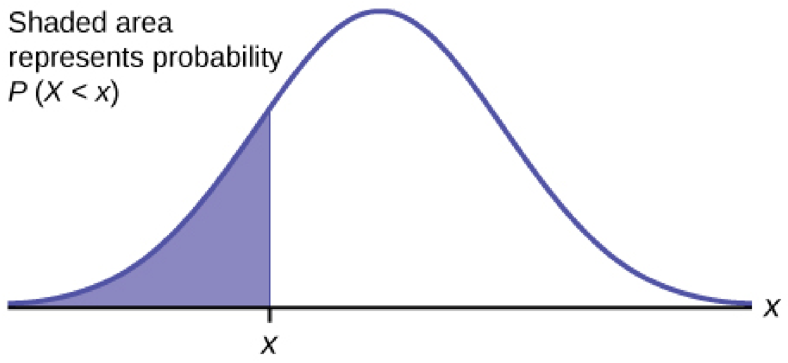
a continuous random variable (RV) $X \sim N(0, 1)$; when X follows the standard normal distribution, it is often noted as $Z \sim N(0, 1)$.

z-score

the linear transformation of the form $z = \frac{x - \mu}{\sigma}$; if this transformation is applied to any normal distribution $X \sim N(\mu, \sigma)$ the result is the standard normal distribution $Z \sim N(0, 1)$. If this transformation is applied to any specific value x of the RV with mean μ and standard deviation σ, the result is called the z-score of x. The z-score allows us to compare data that are normally distributed but scaled differently.

Using the Normal Distribution

The shaded area in the following graph indicates the area to the left of x. This area is represented by the probability $P(X < x)$. Normal tables, computers, and calculators provide or calculate the probability $P(X < x)$.



The area to the right is then $P(X > x) = 1 – P(X < x)$. Remember, $P(X < x)$ = **Area to the left** of the vertical line through x. So $P(X > x) = 1 – P(X < x)$ = **Area to the right** of the vertical line through x.

$P(X < x)$ is the same as $P(X \leq x)$ and $P(X > x)$ is the same as $P(X \geq x)$ for continuous distributions.

Calculations of Probabilities

Probabilities are calculated using technology. There are instructions given as necessary for the TI-83+ and TI-84 calculators.

Note:
NOTE
To calculate the probability, use the probability tables provided in [link] without the use of technology. The tables include instructions for how to use them.

Example:
If the area to the left is 0.0228, then the area to the right is 1 – 0.0228 = 0.9772.

Note:
Try It
Exercise:

Problem:

If the area to the left of x is 0.012, then what is the area to the right?

Solution:

$1 - 0.012 = 0.988$

Example:
The final exam scores in a statistics class were normally distributed with a mean of 63 and a standard deviation of five.

Exercise:

Problem:

a. Find the probability that a randomly selected student scored more than 65 on the exam.

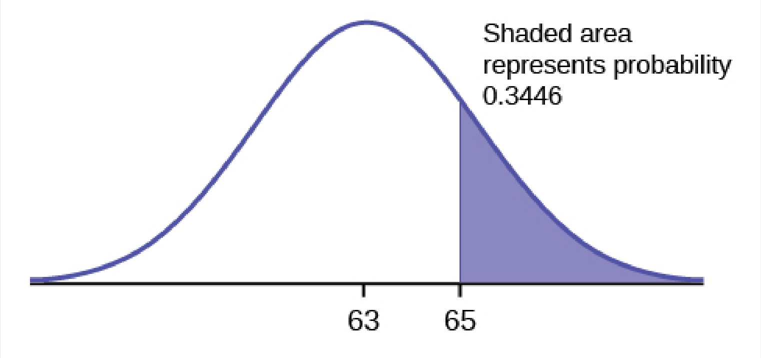
Solution:

a. Let X = a score on the final exam. $X \sim N(63, 5)$, where $\mu = 63$ and $\sigma = 5$

Draw a graph.

Then, find $P(x > 65)$.

$P(x > 65) = 0.3446$



The probability that any student selected at random scores more than 65 is 0.3446.

Note:
Go into `2nd DISTR`.
After pressing `2nd DISTR`, press `2:normalcdf`.
The syntax for the instructions are as follows:
normalcdf(lower value, upper value, mean, standard deviation) For this problem: normalcdf(65,1E99,63,5) = 0.3446. You get 1E99 (= 10^{99}) by pressing `1`, the `EE` key (a 2nd key), and then `99`. The number 10^{99} is way out in the right tail of the normal curve. We are calculating the area between 65 and 10^{99}. In some instances, the lower number of the area might be –1E99 (= -10^{99}). The number -10^{99} is way out in the left tail of the normal curve.

Note:
Historical Note
The TI probability program calculates a z-score and then the probability from the z-score. Before technology, the z-score was looked up in a standard normal probability table (because the math involved is too cumbersome) to find the probability. In this example, a standard normal table with area to the left of the z-score was used.

You calculate the z-score and look up the area to the left. The probability is the area to the right.

$$z = \frac{65 - 63}{5} = 0.4$$

Area to the left is 0.6554.

$P(x > 65) = P(z > 0.4) = 1 - 0.6554 = 0.3446$

Note:
Calculate the z-score:
*Press `2nd Distr`
*Press `3:invNorm`(
*Enter the area to the left of z followed by)
*Press `ENTER`.
For this Example, the steps are
`2nd Distr`
`3:invNorm`(.6554) `ENTER`
The answer is 0.3999 which rounds to 0.4.

Exercise:

Problem:

b. Find the probability that a randomly selected student scored less than 85.

Solution:

b. Draw a graph.

Then find $P(x < 85)$, and shade the graph.

Using a computer or calculator, verify that $P(x < 85) = 1$.

normalcdf(0,85,63,5) = 1 (rounds to one)

The probability that one student scores less than 85 is approximately one (or 100%).

Exercise:

Problem:

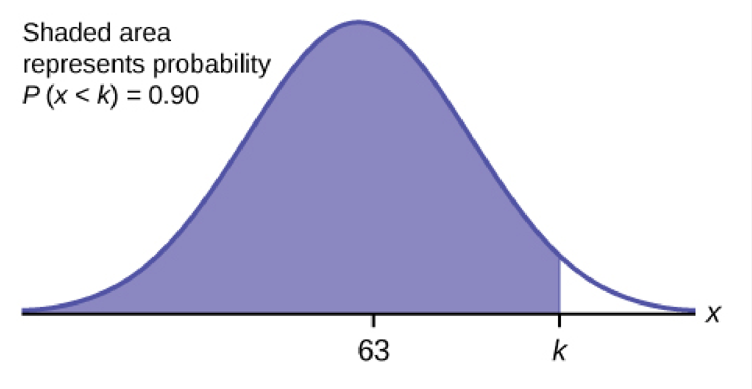
c. Find the 90th percentile (that is, find the score k that has 90% of the scores below k and 10% of the scores above k).

Solution:

c. Find the 90th percentile. For each problem or part of a problem, draw a new graph. Draw the x-axis. Shade the area that corresponds to the 90th percentile.

Let k = the 90th percentile. The variable k is located on the x-axis. $P(x < k)$ is the area to the left of k. The 90th percentile k separates the exam scores into those that are the same or lower than k and those that are the same or higher. Ninety percent of the test scores are the same or lower than k, and ten percent are the same or higher. The variable k is often called a **critical value**.

$k = 69.4$



The 90th percentile is 69.4. This means that 90% of the test scores fall at or below 69.4 and 10% fall at or above. To get this answer on the calculator, follow this step:

Note:
`invNorm` in `2nd DISTR`. invNorm(area to the left, mean, standard deviation)
For this problem, invNorm(0.90,63,5) = 69.4

Exercise:

Problem:

d. Find the 70th percentile (that is, find the score k such that 70% of scores are below k and 30% of the scores are above k).

Solution:

d. Find the 70th percentile.

Draw a new graph and label it appropriately. $k = 65.6$

The 70th percentile is 65.6. This means that 70% of the test scores fall at or below 65.5 and 30% fall at or above.

invNorm(0.70,63,5) = 65.6

Note:
Try It
Exercise:

Problem:

The golf scores for a school team were normally distributed with a mean of 68 and a standard deviation of three.

Find the probability that a randomly selected golfer scored less than 65.

Solution:

normalcdf(-1E99,65,68,3) = 0.1587

Example:

A personal computer is used for office work at home, research, communication, personal finances, education, entertainment, social networking, and a myriad of other things. Suppose that the average number of hours a household personal computer is used for entertainment is two hours per day. Assume the times for entertainment are normally distributed and the standard deviation for the times is half an hour.

Exercise:

Problem:

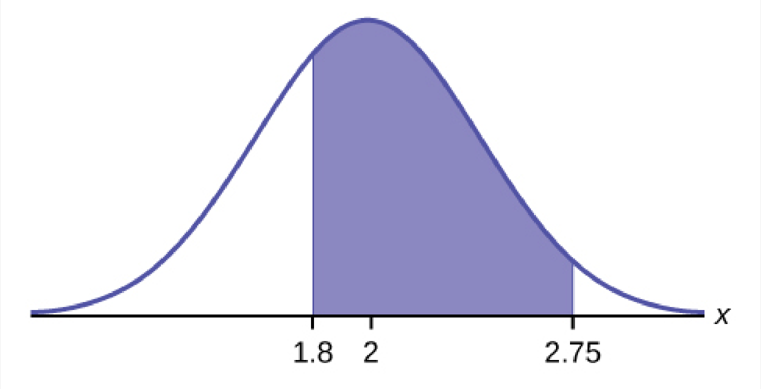
a. Find the probability that a household personal computer is used for entertainment between 1.8 and 2.75 hours per day.

Solution:

a. Let X = the amount of time (in hours) a household personal computer is used for entertainment. $X \sim N(2, 0.5)$ where $\mu = 2$ and $\sigma = 0.5$.

Find $P(1.8 < x < 2.75)$.

The probability for which you are looking is the area **between** x = 1.8 and x = 2.75.
$P(1.8 < x < 2.75) = 0.5886$.



normalcdf(1.8,2.75,2,0.5) = 0.5886

The probability that a household personal computer is used between 1.8 and 2.75 hours per day for entertainment is 0.5886.

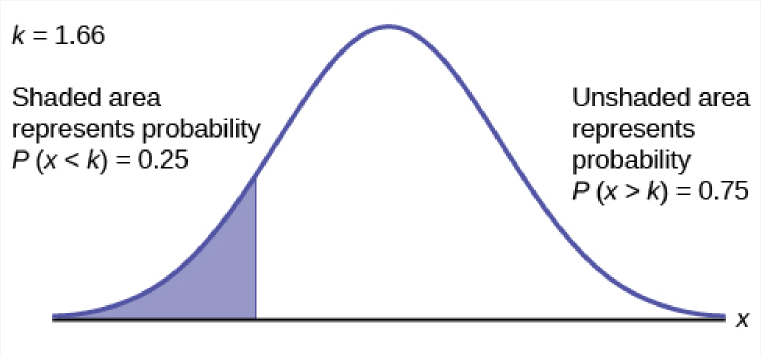
Exercise:

Problem:

b. Find the first quartile for all households that uses a personal computer for entertainment.

Solution:

b. To find the first quartile for all households that uses a personal computer for entertainment, **find the 25th percentile,** k, where $P(x < k) = 0.25$.



invNorm(0.25,2,0.5) = 1.66

The first quartile for all households that uses a personal computer for entertainment is 1.66 hours.

Note:
Try It
Exercise:

Problem:

The golf scores for a school team were normally distributed with a mean of 68 and a standard deviation of three. Find the probability that a golfer scored between 66 and 70.

Solution:

normalcdf(66,70,68,3) = 0.4950

Example:
There are approximately one billion smartphone users in the world today. In the United States the ages 13 to 55+ of smartphone users approximately follow a normal distribution with approximate mean and standard deviation of 36.9 years and 13.9 years, respectively.

Exercise:

Problem:

a. Determine the probability that a random smartphone user in the age range 13 to 55+ is between 23 and 64.7 years old.

Solution:

a. normalcdf(23,64.7,36.9,13.9) = 0.8186

Exercise:

Problem:

b. Determine the probability that a randomly selected smartphone user in the age range 13 to 55+ is at most 50.8 years old.

Solution:

b. normalcdf(–1E99,50.8,36.9,13.9) = 0.8413

Exercise:

Problem:

c. Find the 80th percentile of this distribution, and interpret it in a complete sentence.

Solution:

c.

- invNorm(0.80,36.9,13.9) = 48.6
- The 80th percentile is 48.6 years.
- 80% of the smartphone users in the age range 13 – 55+ are 48.6 years old or less.

Note:

Try It

Use the information in [link] to answer the following questions.

Exercise:

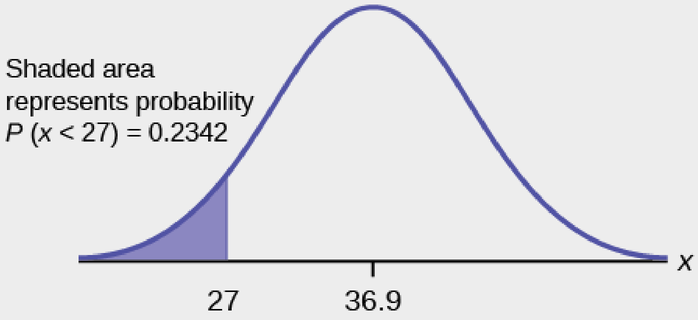
Problem:

a. Find the 30th percentile, and interpret it in a complete sentence.
b. What is the probability that the age of a randomly selected smartphone user in the range 13 to 55+ is less than 27 years old.

Solution:

Let X = a smart phone user whose age is 13 to 55+. $X \sim N(36.9, 13.9)$.

a. To find the 30th percentile, find k such that $P(x < k) = 0.30$.
invNorm(0.30, 36.9, 13.9) = 29.6 years
Thirty percent of smartphone users 13 to 55+ are at most 29.6 years and 70% are at least 29.6 years.
b. Find $P(x < 27)$



normalcdf(0,27,36.9,13.9) = 0.2342
(Note that normalcdf(–1E99,27,36.9,13.9) = 0.2382. The two answers differ only by 0.0040.)

Example:
There are approximately one billion smartphone users in the world today. In the United States the ages 13 to 55+ of smartphone users approximately follow a normal distribution with approximate mean and standard deviation of 36.9 years and 13.9 years respectively. Using this information,

answer the following questions (round answers to one decimal place).

Exercise:

Problem: a. Calculate the interquartile range (*IQR*).

Solution:

a.

- $IQR = Q_3 – Q_1$
- Calculate Q_3 = 75th percentile and Q_1 = 25th percentile.
- invNorm(0.75,36.9,13.9) = Q_3 = 46.2754
- invNorm(0.25,36.9,13.9) = Q_1 = 27.5246
- $IQR = Q_3 – Q_1$ = 18.7508

Exercise:

Problem:

b. Forty percent of the ages that range from 13 to 55+ are at least what age?

Solution:

b.

- Find k where $P(x > k)$ = 0.40 ("At least" translates to "greater than or equal to.")
- 0.40 = the area to the right.
- Area to the left = 1 – 0.40 = 0.60.
- The area to the left of k = 0.60.
- invNorm(0.60,36.9,13.9) = 40.4215.
- k = 40.42.
- Forty percent of the ages that range from 13 to 55+ are at least 40.42 years.

Note:
Try It
Exercise:

Problem:

Two thousand students took an exam. The scores on the exam have an approximate normal distribution with a mean μ = 81 points and standard deviation σ = 15 points.

a. Calculate the first- and third-quartile scores for this exam.
b. The middle 50% of the exam scores are between what two values?

Solution:

a. Q_1 = 25th percentile = invNorm(0.25,81,15) = 70.9
Q_3 = 75th percentile = invNorm(0.75,81,15) = 91.1
b. The middle 50% of the scores are between 70.9 and 91.1.

Example:
A citrus farmer who grows mandarin oranges finds that the diameters of mandarin oranges harvested on his farm follow a normal distribution with a mean diameter of 5.85 cm and a standard deviation of 0.24 cm.

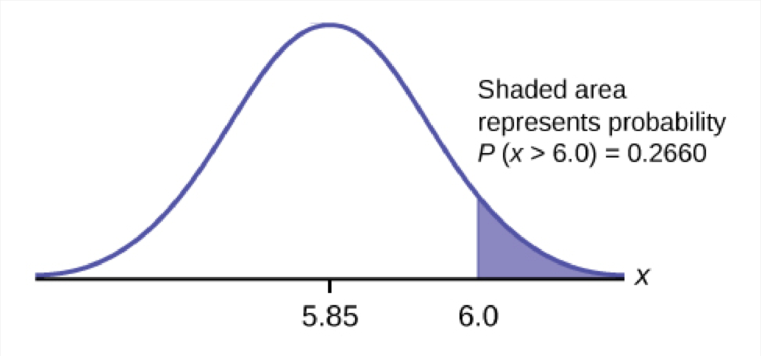
Exercise:

Problem:

a. Find the probability that a randomly selected mandarin orange from this farm has a diameter larger than 6.0 cm. Sketch the graph.

Solution:

a. normalcdf(6,1E99,5.85,0.24) = 0.2660



Exercise:

Problem:

b. The middle 20% of mandarin oranges from this farm have diameters between ______ and ______.

Solution:

b.

- 1 – 0.20 = 0.80
- The tails of the graph of the normal distribution each have an area of 0.40.
- Find *k1*, the 40th percentile, and *k2*, the 60th percentile (0.40 + 0.20 = 0.60).
- *k1* = invNorm(0.40,5.85,0.24) = 5.79 cm
- *k2* = invNorm(0.60,5.85,0.24) = 5.91 cm

Exercise:

Problem:

c. Find the 90th percentile for the diameters of mandarin oranges, and interpret it in a complete sentence.

Solution:

c. 6.16: Ninety percent of the diameter of the mandarin oranges is at most 6.15 cm.

Note:
Try It
Exercise:

Problem: Using the information from [link], answer the following:

a. The middle 40% of mandarin oranges from this farm are between _______ and _______.
b. Find the 16th percentile and interpret it in a complete sentence.

Solution:

a. The middle area = 0.40, so each tail has an area of 0.30.

   1 – 0.40 = 0.60

   The tails of the graph of the normal distribution each have an area of 0.30.

   Find *k1*, the 30th percentile and *k2*, the 70th percentile (0.40 + 0.30 = 0.70).

   k1 = invNorm(0.30,5.85,0.24) = 5.72 cm

$k2$ = invNorm(0.70,5.85,0.24) = 5.98 cm

b. normalcdf(5,1E99,5.85,0.24) = 0.9998

References

“Naegele’s rule.” Wikipedia. Available online at http://en.wikipedia.org/wiki/Naegele's_rule (accessed May 14, 2013).

“403: NUMMI.” Chicago Public Media & Ira Glass, 2013. Available online at http://www.thisamericanlife.org/radio-archives/episode/403/nummi (accessed May 14, 2013).

“Scratch-Off Lottery Ticket Playing Tips.” WinAtTheLottery.com, 2013. Available online at http://www.winatthelottery.com/public/department40.cfm (accessed May 14, 2013).

“Smart Phone Users, By The Numbers.” Visual.ly, 2013. Available online at http://visual.ly/smart-phone-users-numbers (accessed May 14, 2013).

“Facebook Statistics.” Statistics Brain. Available online at http://www.statisticbrain.com/facebook-statistics/(accessed May 14, 2013).

Section Review

The normal distribution, which is continuous, is the most important of all the probability distributions. Its graph is bell-shaped. This bell-shaped curve is used in almost all disciplines. Since it is a continuous distribution, the total area under the curve is one. The parameters of the normal are the mean μ and the standard deviation σ. A special normal distribution, called the standard normal distribution is the distribution of z-scores. Its mean is zero, and its standard deviation is one.

Formula Review

Normal Distribution: $X \sim N(\mu, \sigma)$ where μ is the mean and σ is the standard deviation.

Standard Normal Distribution: $Z \sim N(0, 1)$.

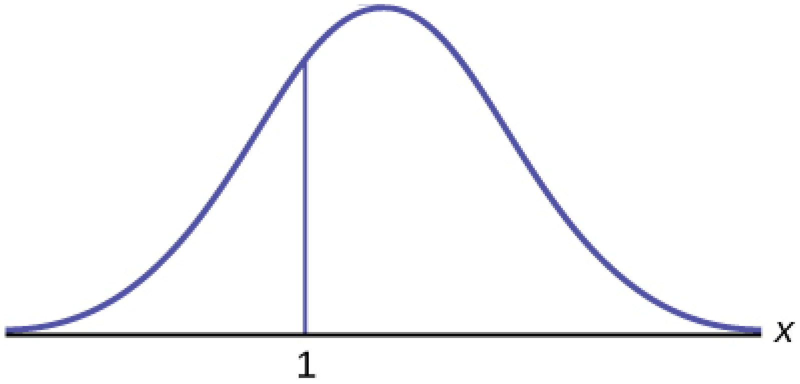
Calculator function for probability: normalcdf (lower x value of the area, upper x value of the area, mean, standard deviation)

Calculator function for the k^{th} percentile: k = invNorm (area to the left of k, mean, standard deviation)

Exercise:

Problem:

How would you represent the area to the left of one in a probability statement?

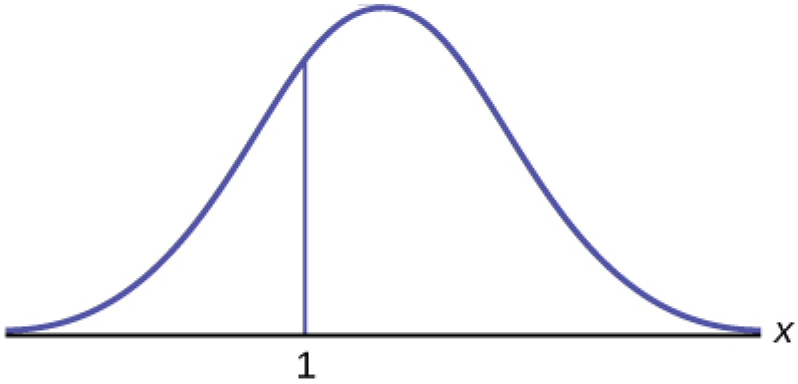


Solution:

$P(x < 1)$

Exercise:

Problem: What is the area to the right of one?



Exercise:

Problem: Is $P(x < 1)$ equal to $P(x \leq 1)$? Why?

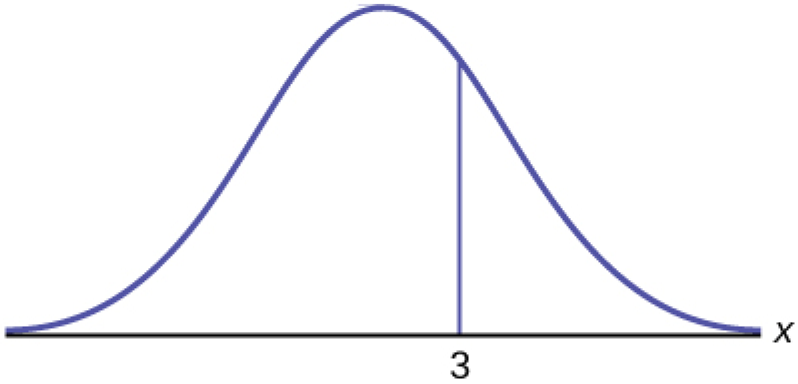
Solution:

Yes, because they are the same in a continuous distribution: $P(x = 1) = 0$.

Exercise:

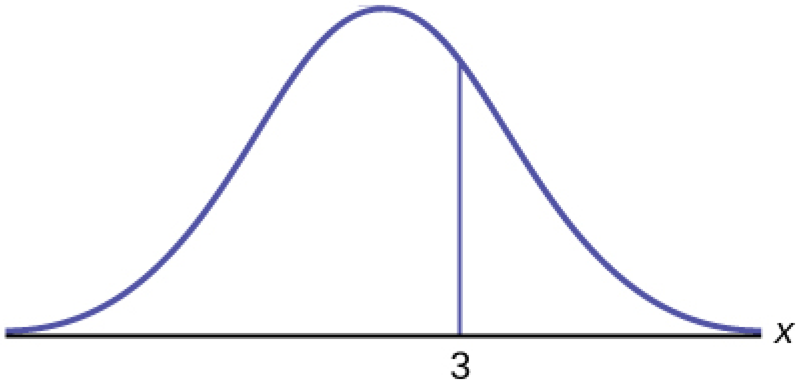
Problem:

How would you represent the area to the left of three in a probability statement?



Exercise:

Problem: What is the area to the right of three?



Solution:

$1 - P(x < 3)$ or $P(x > 3)$.

Exercise:

Problem:

If the area to the left of x in a normal distribution is 0.123, what is the area to the right of x?

Exercise:

Problem:

If the area to the right of x in a normal distribution is 0.543, what is the area to the left of x?

Solution:

$1 - 0.543 = 0.457$

Use the following information to answer the next four exercises:

$X \sim N(54, 8)$

Exercise:

Problem: Find the probability that $x > 56$.

Exercise:

Problem: Find the probability that $x < 30$.

Solution:

0.0013

Exercise:

Problem: Find the 80th percentile.

Exercise:

Problem: Find the 60th percentile.

Solution:

56.03

Exercise:

Problem: $X \sim N(6, 2)$

Find the probability that x is between three and nine.

Exercise:

Problem: $X \sim N(–3, 4)$

Find the probability that x is between one and four.

Solution:

0.1186

Exercise:

Problem: $X \sim N(4, 5)$

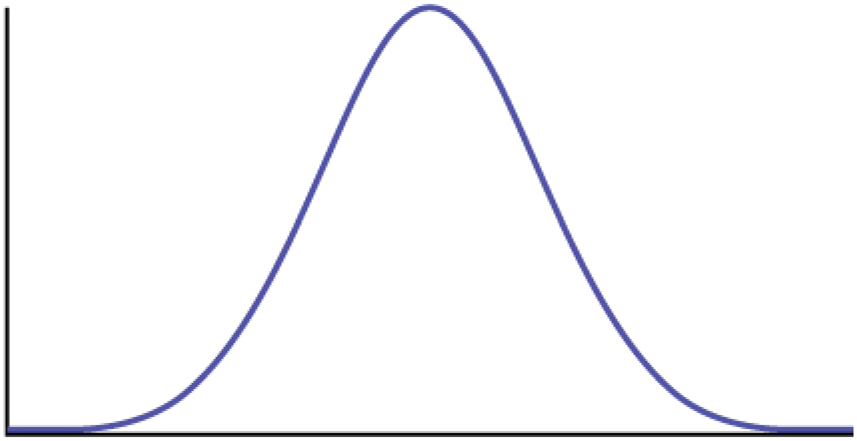
Find the first quartile.

Exercise:

Problem:

Use the following information to answer the next three exercise: The life of Sunshine CD players is normally distributed with a mean of 4.1 years and a standard deviation of 1.3 years. A CD player is guaranteed for three years. We are interested in the length of time a CD player lasts. Find the probability that a CD player will break down during the guarantee period.

a. Sketch the situation. Label and scale the axes. Shade the region corresponding to the probability.



b. $P(0 < x <$ ________________$)$ = _______________ (Use zero for the minimum value of x.)

Solution:

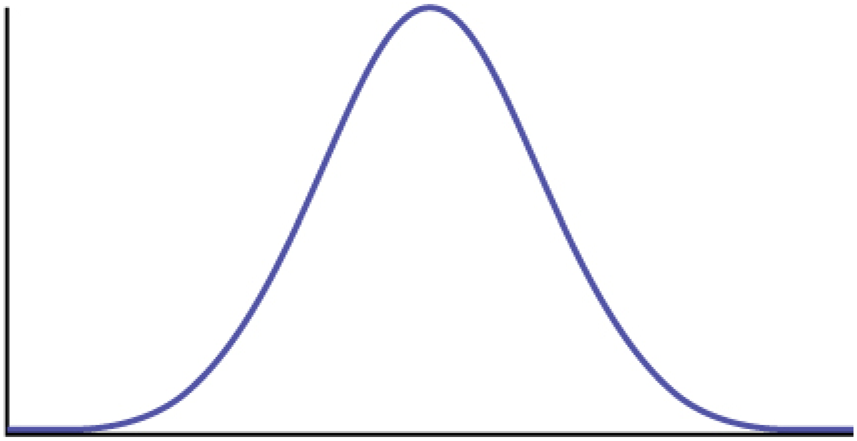
a. Check student's solution.
b. 3, 0.1979

Exercise:

Problem:

Find the probability that a CD player will last between 2.8 and six years.

a. Sketch the situation. Label and scale the axes. Shade the region corresponding to the probability.



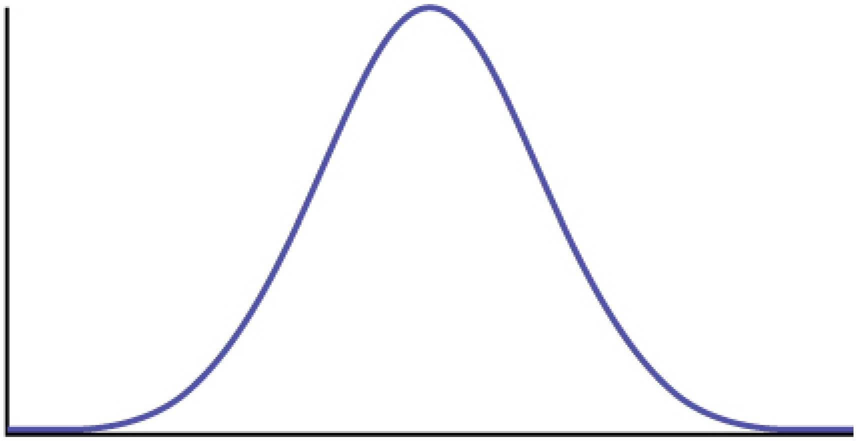
b. $P($___________ $< x <$ ___________$) =$ ___________

Exercise:

Problem:

Find the 70th percentile of the distribution for the time a CD player lasts.

a. Sketch the situation. Label and scale the axes. Shade the region corresponding to the lower 70%.



b. $P(x < k)$ = __________ Therefore, k = _________

Solution:

a. Check student’s solution.
b. 0.70, 4.78 years

Homework

Use the following information to answer the next two exercises: The patient recovery time from a particular surgical procedure is normally distributed with a mean of 5.3 days and a standard deviation of 2.1 days.

Exercise:

Problem:

What is the probability of spending more than two days in recovery?

a. 0.0580
b. 0.8447
c. 0.0553
d. 0.9420

Exercise:

Problem: The 90th percentile for recovery times is?

a. 8.89
b. 7.07
c. 7.99
d. 4.32

Solution:

c

Use the following information to answer the next three exercises: The length of time it takes to find a parking space at 9 A.M. follows a normal distribution with a mean of five minutes and a standard deviation of two minutes.

Exercise:

Problem:

Based upon the given information and numerically justified, would you be surprised if it took less than one minute to find a parking space?

a. Yes
b. No
c. Unable to determine

Exercise:

Problem:

Find the probability that it takes at least eight minutes to find a parking space.

a. 0.0001
b. 0.9270

c. 0.1862
d. 0.0668

Solution:

d

Exercise:

Problem:

Seventy percent of the time, it takes more than how many minutes to find a parking space?

a. 1.24
b. 2.41
c. 3.95
d. 6.05

Exercise:

Problem:

According to a study done by De Anza students, the height for Asian adult males is normally distributed with an average of 66 inches and a standard deviation of 2.5 inches. Suppose one Asian adult male is randomly chosen. Let X = height of the individual.

a. X ~ _____(_____,_____)
b. Find the probability that the person is between 65 and 69 inches. Include a sketch of the graph, and write a probability statement.
c. Would you expect to meet many Asian adult males over 72 inches? Explain why or why not, and justify your answer numerically.
d. The middle 40% of heights fall between what two values? Sketch the graph, and write the probability statement.

Solution:

a. $X \sim N(66, 2.5)$
b. 0.5404
c. No, the probability that an Asian male is over 72 inches tall is 0.0082

Exercise:

Problem:

IQ is normally distributed with a mean of 100 and a standard deviation of 15. Suppose one individual is randomly chosen. Let X = IQ of an individual.

a. X ~ _____(_____,_____)
b. Find the probability that the person has an IQ greater than 120. Include a sketch of the graph, and write a probability statement.
c. MENSA is an organization whose members have the top 2% of all IQs. Find the minimum IQ needed to qualify for the MENSA organization. Sketch the graph, and write the probability statement.
d. The middle 50% of IQs fall between what two values? Sketch the graph and write the probability statement.

Exercise:

Problem:

The percent of fat calories that a person in America consumes each day is normally distributed with a mean of about 36 and a standard deviation of 10. Suppose that one individual is randomly chosen. Let X = percent of fat calories.

a. X ~ _____(_____,_____)
b. Find the probability that the percent of fat calories a person consumes is more than 40. Graph the situation. Shade in the area to be determined.

c. Find the first quartile of percent of fat calories. Sketch the graph and write the probability statement.

Solution:

a. $X \sim N(36, 10)$
b. The probability that a person consumes more than 40% of their calories as fat is 0.3446.
c. Approximately 25% of people consume less than 29.26% of their calories as fat.

Exercise:

Problem:

Suppose that the distance of fly balls hit to the outfield (in baseball) is normally distributed with a mean of 250 feet and a standard deviation of 50 feet.

a. If X = distance in feet for a fly ball, then X ~ _____(_____,_____)
b. If one fly ball is randomly chosen from this distribution, what is the probability that this ball traveled fewer than 220 feet? Sketch the graph. Scale the horizontal axis X. Shade the region corresponding to the probability. Find the probability.
c. Find the 80th percentile of the distribution of fly balls. Sketch the graph, and write the probability statement.

Exercise:

Problem:

In China, four-year-olds average three hours a day unsupervised. Most of the unsupervised children live in rural areas, considered safe. Suppose that the standard deviation is 1.5 hours and the amount of time spent alone is normally distributed. We randomly select one Chinese four-year-old living in a rural area. We are interested in the amount of time the child spends alone per day.

a. In words, define the random variable X.
b. X ~ _____(_____,_____)
c. Find the probability that the child spends less than one hour per day unsupervised. Sketch the graph, and write the probability statement.
d. What percent of the children spend over ten hours per day unsupervised?
e. Seventy percent of the children spend at least how long per day unsupervised?

Solution:

a. X = number of hours that a Chinese four-year-old in a rural area is unsupervised during the day.
b. X ~ $N(3, 1.5)$
c. The probability that the child spends less than one hour a day unsupervised is 0.0912.
d. The probability that a child spends over ten hours a day unsupervised is less than 0.0001.
e. 2.21 hours

Exercise:

Problem:

In the 1992 presidential election, Alaska’s 40 election districts averaged 1,956.8 votes per district for President Clinton. The standard deviation was 572.3. (There are only 40 election districts in Alaska.) The distribution of the votes per district for President Clinton was bell-shaped. Let X = number of votes for President Clinton for an election district.

a. State the approximate distribution of X.
b. Is 1,956.8 a population mean or a sample mean? How do you know?
c. Find the probability that a randomly selected district had fewer than 1,600 votes for President Clinton. Sketch the graph and write the probability statement.
d. Find the probability that a randomly selected district had between 1,800 and 2,000 votes for President Clinton.
e. Find the third quartile for votes for President Clinton.

Exercise:

Problem:

Suppose that the duration of a particular type of criminal trial is known to be normally distributed with a mean of 21 days and a standard deviation of seven days.

a. In words, define the random variable X.
b. X ~ _____(_____,_____)
c. If one of the trials is randomly chosen, find the probability that it lasted at least 24 days. Sketch the graph and write the probability statement.
d. Sixty percent of all trials of this type are completed within how many days?

Solution:

a. X = the distribution of the number of days a particular type of criminal trial will take
b. $X \sim N(21, 7)$
c. The probability that a randomly selected trial will last more than 24 days is 0.3336.
d. 22.77

Exercise:

Problem:

Terri Vogel, an amateur motorcycle racer, averages 129.71 seconds per 2.5 mile lap (in a seven-lap race) with a standard deviation of 2.28 seconds. The distribution of her race times is normally distributed. We are interested in one of her randomly selected laps.

a. In words, define the random variable X.
b. X ~ _____(_____,_____)
c. Find the percent of her laps that are completed in less than 130 seconds.
d. The fastest 3% of her laps are under _____.
e. The middle 80% of her laps are from _______ seconds to _______ seconds.

Exercise:

Problem:

Thuy Dau, Ngoc Bui, Sam Su, and Lan Voung conducted a survey as to how long customers at Lucky claimed to wait in the checkout line until their turn. Let X = time in line. The following table displays the ordered real data (in minutes):

0.50	4.25	5	6	7.25
1.75	4.25	5.25	6	7.25
2	4.25	5.25	6.25	7.25
2.25	4.25	5.5	6.25	7.75
2.25	4.5	5.5	6.5	8
2.5	4.75	5.5	6.5	8.25
2.75	4.75	5.75	6.5	9.5
3.25	4.75	5.75	6.75	9.5
3.75	5	6	6.75	9.75
3.75	5	6	6.75	10.75

a. Calculate the sample mean and the sample standard deviation.
b. Construct a histogram.
c. Draw a smooth curve through the midpoints of the tops of the bars.
d. In words, describe the shape of your histogram and smooth curve.
e. Let the sample mean approximate μ and the sample standard deviation approximate σ. The distribution of X can then be approximated by X ~ _____(_____,_____)
f. Use the distribution in part e to calculate the probability that a person will wait fewer than 6.1 minutes.
g. Determine the cumulative relative frequency for waiting less than 6.1 minutes.
h. Why aren’t the answers to part f and part g exactly the same?
i. Why are the answers to part f and part g as close as they are?
j. If only ten customers has been surveyed rather than 50, do you think the answers to part f and part g would have been closer together or farther apart? Explain your conclusion.

Solution:

a. mean = 5.51, s = 2.15
b. Check student's solution.
c. Check student's solution.
d. Check student's solution.
e. $X \sim N(5.51, 2.15)$
f. 0.6029
g. The cumulative frequency for less than 6.1 minutes is 0.64.
h. The answers to part f and part g are not exactly the same, because the normal distribution is only an approximation to the real one.
i. The answers to part f and part g are close, because a normal distribution is an excellent approximation when the sample size is greater than 30.
j. The approximation would have been less accurate, because the smaller sample size means that the data does not fit normal curve as well.

Exercise:

Problem:

Suppose that Ricardo and Anita attend different colleges. Ricardo’s GPA is the same as the average GPA at his school. Anita’s GPA is 0.70 standard deviations above her school average. In complete sentences, explain why each of the following statements may be false.

a. Ricardo’s actual GPA is lower than Anita’s actual GPA.
b. Ricardo is not passing because his z-score is zero.
c. Anita is in the 70th percentile of students at her college.

Exercise:

Problem:

The following table shows a sample of the maximum capacity (maximum number of spectators) of sports stadiums. The table does not include horse-racing or motor-racing stadiums.

40,000	40,000	45,050	45,500	46,249	48,134
49,133	50,071	50,096	50,466	50,832	51,100
51,500	51,900	52,000	52,132	52,200	52,530
52,692	53,864	54,000	55,000	55,000	55,000
55,000	55,000	55,000	55,082	57,000	58,008
59,680	60,000	60,000	60,492	60,580	62,380
62,872	64,035	65,000	65,050	65,647	66,000
66,161	67,428	68,349	68,976	69,372	70,107
70,585	71,594	72,000	72,922	73,379	74,500
75,025	76,212	78,000	80,000	80,000	82,300

a. Calculate the sample mean and the sample standard deviation for the maximum capacity of sports stadiums (the data).
b. Construct a histogram.
c. Draw a smooth curve through the midpoints of the tops of the bars of the histogram.
d. In words, describe the shape of your histogram and smooth curve.

e. Let the sample mean approximate μ and the sample standard deviation approximate σ. The distribution of X can then be approximated by X ~ _____(_____,_____).
f. Use the distribution in part e to calculate the probability that the maximum capacity of sports stadiums is less than 67,000 spectators.
g. Determine the cumulative relative frequency that the maximum capacity of sports stadiums is less than 67,000 spectators. Hint: Order the data and count the sports stadiums that have a maximum capacity less than 67,000. Divide by the total number of sports stadiums in the sample.
h. Why aren't the answers to part f and part g exactly the same?

Solution:

1. mean = 60,136
   s = 10,468
2. Answers will vary.
3. Answers will vary.
4. Answers will vary.
5. $X \sim N(60136, 10468)$
6. 0.7440
7. The cumulative relative frequency is 43/60 = 0.717.
8. The answers for part f and part g are not the same, because the normal distribution is only an approximation.

Exercise:

Problem:

An expert witness for a paternity lawsuit testifies that the length of a pregnancy is normally distributed with a mean of 280 days and a standard deviation of 13 days. An alleged father was out of the country from 240 to 306 days before the birth of the child, so the pregnancy would have been less than 240 days or more than 306 days long if he was the father. The birth was uncomplicated, and the child needed no medical intervention. What is the probability that he was NOT the father? What is the probability that he could be the father? Calculate the z-scores first, and then use those to calculate the probability.

Exercise:

Problem:

A NUMMI assembly line, which has been operating since 1984, has built an average of 6,000 cars and trucks a week. Generally, 10% of the cars were defective coming off the assembly line. Suppose we draw a random sample of n = 100 cars. Let X represent the number of defective cars in the sample. What can we say about X in regard to the 68-95-99.7 empirical rule (one standard deviation, two standard deviations and three standard deviations from the mean are being referred to)? Assume a normal distribution for the defective cars in the sample.

Solution:

- n = 100; p = 0.1; q = 0.9
- μ = np = (100)(0.10) = 10
- σ = $\sqrt{npq}$ = $\sqrt{(100)(0.1)(0.9)}$ = 3

i. z = ±1: x_1 = μ + $z\sigma$ = 10 + 1(3) = 13 and x_2 = μ – $z\sigma$ = 10 – 1(3) = 7. Thus, 68% of the defective cars will fall between seven and 13.
ii. z = ±2: x_1 = μ + $z\sigma$ = 10 + 2(3) = 16 and x_2 = μ – $z\sigma$ = 10 – 2(3) = 4. Thus, 95% of the defective cars will fall between four and 16

iii. $z = \pm 3$: $x_1 = \mu + z\sigma = 10 + 3(3) = 19$ and $x_2 = \mu - z\sigma = 10 - 3(3) = 1$. Thus, 99.7% of the defective cars will fall between one and 19.

Exercise:

Problem:

We flip a coin 100 times ($n = 100$) and note that it only comes up heads 20% ($p = 0.20$) of the time. The distribution for the number of times the coin lands on heads is approximately symmetric and bell-shaped with a mean and standard deviation of $\mu = 20$ and $\sigma = 4$ (verify the mean and standard deviation). Solve the following, using the Empirical Rule:

a. There is about a 68% chance that the number of heads will be somewhere between ___ and ___.
b. There is about a ____chance that the number of heads will be somewhere between 12 and 28.
c. There is about a ____ chance that the number of heads will be somewhere between eight and 32.

Exercise:

Problem:

A $1 scratch off lotto ticket will be a winner one out of five times. Out of a shipment of $n = 190$ lotto tickets, find the probability for the lotto tickets that there are

a. somewhere between 34 and 54 prizes.
b. somewhere between 54 and 64 prizes.
c. more than 64 prizes.

Solution:

- $n = 190$; $p = \frac{1}{5} = 0.2$; $q = 0.8$

- $\mu = np$ = (190)(0.2) = 38
- $\sigma = \sqrt{npq} = \sqrt{(190)(0.2)(0.8)}$ = 5.5136

a. For this problem: $P(34 < x < 54)$ = normalcdf(34,54,38,5.5136) = 0.7641
b. For this problem: $P(54 < x < 64)$ = normalcdf(54,64,38,5.5136) = 0.0019
c. For this problem: $P(x > 64)$ = normalcdf(64,1E99,38,5.5136) = 0.0000012 (approximately 0)

Exercise:

Problem:

Facebook provides a variety of statistics on its Web site that detail the growth and popularity of the site.

On average, 28 percent of 18 to 34 year olds check their Facebook profiles before getting out of bed in the morning. Suppose this percentage follows a normal distribution with a standard deviation of five percent.

a. Find the probability that the percent of 18 to 34-year-olds who check Facebook before getting out of bed in the morning is at least 30.
b. Find the 95th percentile, and express it in a sentence.

The Central Limit Theorem: Introduction class="introduction"

If you want to figure out the distribution of the change people carry in their pockets, using the central limit theorem and assuming your sample is large enough, you will find that the distribution is normal and bell-shaped. (credit: John Lodder)

Note:
Chapter Objectives
By the end of this chapter, the student should be able to:

- Recognize central limit theorem problems.
- Classify continuous word problems by their distributions.
- Apply and interpret the central limit theorem for means.

Why are we so concerned with means? Two reasons are: they give us a middle ground for comparison, and they are easy to calculate. In this chapter, you will study means and the **central limit theorem**.

The **central limit theorem** (clt for short) is one of the most powerful and useful ideas in all of statistics. This theorem is concerned with drawing finite samples size n from a population with a known mean, μ, and a known standard deviation, σ. The theorem says that if we collect samples of size n with a "large enough n," calculate each sample's mean, and create a histogram of those means, then the resulting histogram will tend to have an approximate normal bell shape.

The size of the sample, n, that is required in order to be "large enough" depends on the original population from which the samples are drawn (the sample size should be at least 30 or the data should come from a normal distribution). If the original population is far from normal, then more observations are needed for the sample means to be normal. **Sampling is done with replacement.**

It would be difficult to overstate the importance of the central limit theorem in statistical theory. Knowing that data, even if its distribution is not normal, behaves in a predictable way is a powerful tool.

Note:

Collaborative Classroom Activity

Suppose eight of you roll one fair die ten times, seven of you roll two fair dice ten times, nine of you roll five fair dice ten times, and 11 of you roll ten fair dice ten times.

Each time a person rolls more than one die, he or she calculates the sample **mean** of the faces showing. For example, one person might roll five fair dice and get 2, 2, 3, 4, 6 on one roll.

The mean is $\frac{2+2+3+4+6}{5}$ = 3.4. The 3.4 is one mean when five fair dice are rolled. This same person would roll the five dice nine more times and calculate nine more means for a total of ten means.

Your instructor will pass out the dice to several people. Roll your dice ten times. For each roll, record the faces, and find the mean. Round to the nearest 0.5.

Your instructor (and possibly you) will produce one graph (it might be a histogram) for one die, one graph for two dice, one graph for five dice, and one graph for ten dice. Since the "mean" when you roll one die is just the face on the die, what distribution do these **means** appear to be representing?

Draw the graph for the means using two dice. Do the sample means show any kind of pattern?

Draw the graph for the means using five dice. Do you see any pattern emerging?

Finally, draw the graph for the means using ten dice. Do you see any pattern to the graph? What can you conclude as you increase the number of dice?

As the number of dice rolled increases from one to two to five to ten, the following is happening:

1. The mean of the sample means remains approximately the same.
2. The spread of the sample means (the standard deviation of the sample means) gets smaller.
3. The graph appears steeper and thinner.

You have just demonstrated the central limit theorem (clt).

The central limit theorem tells you that as you increase the number of dice, **the distribution of the sample means tends toward a normal distribution.**

Glossary

Sampling Distribution

Given simple random samples of size n from a given population with a measured characteristic such as mean, proportion, or standard deviation for each sample, the probability distribution of all the measured characteristics is called a sampling distribution.

The Central Limit Theorem for Sample Means (Averages)

Suppose X is a random variable with a distribution that may be known or unknown (it can be any distribution) and

- μ_x = the mean of X
- σ_x = the standard deviation of X

If you draw random samples of size n, then the distribution of the random variable X, which represents the mean of a random sample, is called the **sampling distribution of the sample mean**.

In other words, the **sampling distribution of the sample mean** is the distribution of every possible sample mean that can be obtained from selecting a random sample of size n from the original population (the distribution of X).

The mean of the sampling distribution of the sample mean will always be the same as the mean of the original population.

Equation:

$$\mu_x = \mu_x$$

Note that this is true regardless of the sample size.

The standard deviation of the sampling distribution of the sample mean is equal to the standard deviation of the original population divided by the square root of the sample size:

Equation:

$$\sigma_x = \frac{\sigma_x}{\sqrt{n}}$$

If the random variable X has a normal distribution, then the sampling distribution of the sample mean will also have a normal distribution, regardless of the sample size.

In cases where the original population is not normal or the population distribution is unknown, which is most likely the case, the central theorem can be quite helpful.

The Central Limit Theorem

The **central limit theorem** for sample means says that **as the sample size increases**, the sampling distribution of the sample mean grows closer to a **normal distribution**, regardless of the shape of the original population distribution.

(Generally, a good rule of thumb is to use a sample size of at least 30, to ensure a sampling distribution that will be approximately normal. Unless of course the original population is known to be normal, in which case the sampling distribution of the sample mean will be guaranteed to normal.)

To sum things up, if X is a random variable with mean μ_x and standard deviation σ_x and either X is normally distributed or $n \geq 30$, then

Equation:

$$X \sim N\left(\mu_x, \frac{\sigma_x}{\sqrt{n}}\right)$$

The random variable X has a different z-score associated with it from that of the random variable X. The mean x is the value of X in one sample. Plugging in the mean and standard deviation of the sampling distribution of the sample mean into the z-score formula, we obtain the following formula.

Equation:

$$z_x = \frac{x - \mu_x}{\left(\frac{\sigma_x}{\sqrt{n}}\right)}$$

μ_x is the average of both X and X.

$\sigma_x = \frac{\sigma_x}{\sqrt{n}}$ = standard deviation of X and is called the **standard error of the mean.**

Note:
To find probabilities for means on the calculator, follow these steps.
2nd DISTR
2:normalcdf

$$normalcdf\left(lower\ value,\ upper\ value,\ mean,\ \frac{standard\ deviation}{\sqrt{sample\ size}}\right)$$

where:

- $mean$ is the mean of the original distribution
- $standard\ deviation$ is the standard deviation of the original distribution
- $sample\ size = n$

Example:
An unknown distribution has a mean of 90 and a standard deviation of 15. A sample of size n = 25 is drawn randomly from the population.

Exercise:

Problem:

a. Find the probability that the **sample mean** is between 85 and 92.

Solution:

a. Let X = one value from the original unknown population. The probability question asks you to find a probability for the **sample mean**.

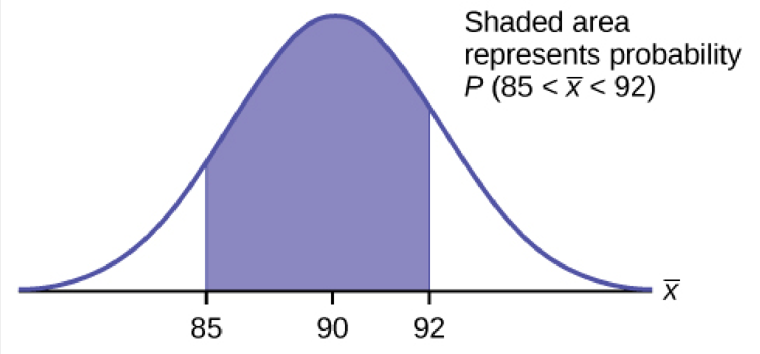
Let X = the mean of a sample of size 25. Since μ_x = 90, σ_x = 15, and n = 25,

$X \sim N\left(90, \frac{15}{\sqrt{25}}\right)$.

Find $P(85 < x < 92)$. Draw a graph.

$P(85 < x < 92)$ = 0.6997

The probability that the sample mean is between 85 and 92 is 0.6997.



Note:
`normalcdf`(lower value, upper value, mean, standard error of the mean)
The parameter list is abbreviated (lower value, upper value, μ, $\frac{\sigma}{\sqrt{n}}$)
`normalcdf`(85,92,90,$\frac{15}{\sqrt{25}}$) = 0.6997

Exercise:

Problem:

b. Find the value of the sample mean that is two standard deviations above the expected value, 90.

Solution:

b. To find the value that is two standard deviations above the expected value 90, use the z-score formula:

$$z_x = \frac{x - \mu_x}{\left(\frac{\sigma_x}{\sqrt{n}}\right)}$$

where the number of standard deviations, z_x, is 2, the expected value, μ_x, is 90, the standard deviation of the original distribution, σ_x is 15, and the sample size n is 25.

Plugging in the known values and solving for x
Equation:

$$2 = \frac{x-90}{\left(15/\sqrt{25}\right)}$$
$$2 = \frac{x-90}{3}$$
$$6 = x - 90$$
$$x = 96$$

The value of the sample mean that is two standard deviations above the expected value is 96.

The standard error of the mean is $\frac{\sigma_x}{\sqrt{n}}$ = $\frac{15}{\sqrt{25}}$ = 3. Recall that the standard error of the mean is a description of how far (on average) that the sample mean will be from the population mean in repeated simple random samples of size n.

Note:
Try It
Exercise:

Problem:

An unknown distribution has a mean of 45 and a standard deviation of eight. Samples of size n = 30 are drawn randomly from the population. Find the probability that the sample mean is between 42 and 50.

Solution:

$P(42 < x < 50)$ = $\left(42{,}50{,}45,\frac{8}{\sqrt{30}}\right)$ = 0.9797

Example:
Exercise:

Problem:

The length of time, in hours, it takes an "over 40" group of people to play one soccer match is normally distributed with a **mean of two hours** and a **standard deviation of 0.5 hours**. A **sample of size *n* = 50** is drawn randomly from the population. Find the probability that the **sample mean** is between 1.8 hours and 2.3 hours.

Solution:

Let X = the time, in hours, it takes to play one soccer match.

The probability question asks you to find a probability for the **sample mean time**, in hours, it takes to play one soccer match.

Let X = the **mean** time, in hours, it takes to play one soccer match.

If μ_x = __________, σ_x = ___________, and n = ____________, then X ~ N(_______, _______) by the **central limit theorem for means**.

$\mu_x = 2$, $\sigma_x = 0.5$, $n = 50$, and $X \sim N\left(2, \frac{0.5}{\sqrt{50}}\right)$

Find $P(1.8 < x < 2.3)$. Draw a graph.

$P(1.8 < x < 2.3) = 0.9977$

`normalcdf`$\left(1.8,2.3,2,\frac{.5}{\sqrt{50}}\right) = 0.9977$

The probability that the mean time is between 1.8 hours and 2.3 hours is 0.9977.

Note:
Try It
Exercise:

Problem:

The length of time taken on the SAT for a group of students is normally distributed with a mean of 2.5 hours and a standard deviation of 0.25 hours. A sample size of $n = 60$ is drawn randomly from the population. Find the probability that the sample mean is between two hours and three hours.

Solution:

$P(2 < x < 3) = \text{normalcdf}\left(2, 3, 2.5, \frac{0.25}{\sqrt{60}}\right) = 1$

Note:
To find percentiles for means on the calculator, follow these steps.
2nd DIStR
3:invNorm

k = invNorm$\left(\text{area to the left of } k\text{, mean, } \frac{\mathit{standard\ deviation}}{\sqrt{\mathit{sample\ size}}}\right)$

where:

- k = the k^{th} percentile
- mean is the mean of the original distribution
- $\mathit{standard\ deviation}$ is the standard deviation of the original distribution
- $\mathit{sample\ size}$ = n

Example:
Exercise:

Problem:

In a recent study reported Oct. 29, 2012 on the Flurry Blog, the mean age of tablet users is 34 years. Suppose the standard deviation is 15 years. Take a sample of size n = 100.

a. What are the mean and standard deviation for the sample mean ages of tablet users?
b. What does the distribution look like?
c. Find the probability that the sample mean age is more than 30 years (the reported mean age of tablet users in this particular study).
d. Find the 95th percentile for the sample mean age (to one decimal place).

Solution:

a. Since the sample mean tends to target the population mean, we have $\mu_x = \mu_x$ = 34. The standard deviation of the sample means is given by $\sigma_x = \frac{\sigma_x}{\sqrt{n}}$ = $\frac{15}{\sqrt{100}}$ = $\frac{15}{10}$ = 1.5

b. The central limit theorem states that for large sample sizes (n), the sampling distribution will be approximately normal.
c. The probability that the sample mean age is more than 30 is given by $P(x > 30)$ = `normalcdf`(30,1E99,34,1.5) = 0.9962
d. Let k = the 95th percentile.
k = invNorm$\left(0.95,34,\frac{15}{\sqrt{100}}\right)$ = 36.5

Note:
Try It
Exercise:

Problem:

In an article on Flurry Blog, a gaming marketing gap for men between the ages of 30 and 40 is identified. You are researching a startup game targeted at the 35-year-old demographic. Your idea is to develop a strategy game that can be played by men from their late 20s through their late 30s. Based on the article's data, industry research shows that the average strategy player is 28 years old with a standard deviation of 4.8 years. You take a sample of 100 randomly selected gamers. If your target market is 29- to 35-year-olds, should you continue with your development strategy?

Solution:

You need to determine the probability for men whose mean age is between 29 and 35 years of age wanting to play a strategy game.

$P(29 < x < 35)$ = `normalcdf`$\left(29,35,28,\frac{4.8}{\sqrt{100}}\right)$ = 0.0186

You can conclude there is approximately a 1.9% chance that your game will be played by men whose mean age is between 29 and 35.

Example:
Exercise:

Problem:

The mean number of minutes for app engagement by a tablet user is 8.2 minutes. Suppose the standard deviation is one minute. Take a sample of 60.

a. What are the mean and standard deviation for the sample mean number of app engagement by a tablet user?
b. What is the standard error of the mean?
c. Find the 90th percentile for the sample mean time for app engagement for a tablet user. Interpret this value in a complete sentence.
d. Find the probability that the sample mean is between eight minutes and 8.5 minutes.

Solution:

a. $\mu_x = \mu = 8.2;\ \sigma_x = \frac{\sigma}{\sqrt{n}} = \frac{1}{\sqrt{60}} = 0.13$
b. The standard error of the mean is another name for the standard deviation of the sample mean and its value is 0.13.
c. Let k = the 90th percentile
k = `invNorm`$\left(0.90,8.2,\frac{1}{\sqrt{60}}\right)$ = 8.37. This values indicates that 90 percent of the average app engagement time for table users is less than 8.37 minutes.
d. $P(8 < x < 8.5)$ = `normalcdf`$\left(8,8.5,8.2,\frac{1}{\sqrt{60}}\right)$ = 0.9293

Note:
Try It
Exercise:

Problem:

Cans of a cola beverage claim to contain 16 ounces. The amounts in a sample are measured and the statistics are n = 34, x = 16.01 ounces. If the cans are filled so that μ = 16.00 ounces (as labeled) and σ = 0.143 ounces, find the probability that a sample of 34 cans will have an average amount greater than 16.01 ounces. Do the results suggest that cans are filled with an amount greater than 16 ounces?

Solution:

We have $P(x > 16.01)$ = `normalcdf`$\left(16.01,1E99,16,\frac{0.143}{\sqrt{34}}\right)$ = 0.3417. Since there is a 34.17% probability that the average sample weight is greater than 16.01 ounces, we should be skeptical of the company’s claimed volume. If I am a consumer, I should be glad that I am probably receiving free cola. If I am the manufacturer, I need to determine if my bottling processes are outside of acceptable limits.

References

Baran, Daya. “20 Percent of Americans Have Never Used Email.”WebGuild, 2010. Available online at http://www.webguild.org/20080519/20-percent-of-americans-have-never-used-email (accessed May 17, 2013).

Data from The Flurry Blog, 2013. Available online at http://blog.flurry.com (accessed May 17, 2013).

Data from the United States Department of Agriculture.

Section Review

In a population whose distribution may be known or unknown, if the size (n) of samples is sufficiently large, the distribution of the sample means will be approximately normal. The mean of the sample means will equal the population mean. The standard deviation of the distribution of the sample means, called the standard error of the mean, is equal to the population standard deviation divided by the square root of the sample size (n).

Formula Review

The Central Limit Theorem for Sample Means:

As n becomes large (at least 30 is a good rule of thumb) $X \sim N\left(\mu_x, \frac{\sigma_x}{\sqrt{n}}\right)$.

Mean of X: μ_x

Standard Error of the Mean (Standard Deviation of X): $\frac{\sigma_x}{\sqrt{n}}$

z-score of the sample mean: $z_x = \frac{x-\mu_x}{\left(\frac{\sigma_x}{\sqrt{n}}\right)}$

Use the following information to answer the next six exercises: Yoonie is a personnel manager in a large corporation. Each month she must review 16 of the employees. From past experience, she has found that the reviews take her approximately four hours each to do with a population standard deviation of 1.2 hours. Let X be the random variable representing the time it takes her to complete one review. Assume X is normally distributed. Let X be the random variable representing the mean time to complete the 16 reviews. Assume that the 16 reviews represent a random set of reviews.

Exercise:

Problem: What is the mean, standard deviation, and sample size?

Solution:

mean = 4 hours; standard deviation = 1.2 hours; sample size = 16

Exercise:

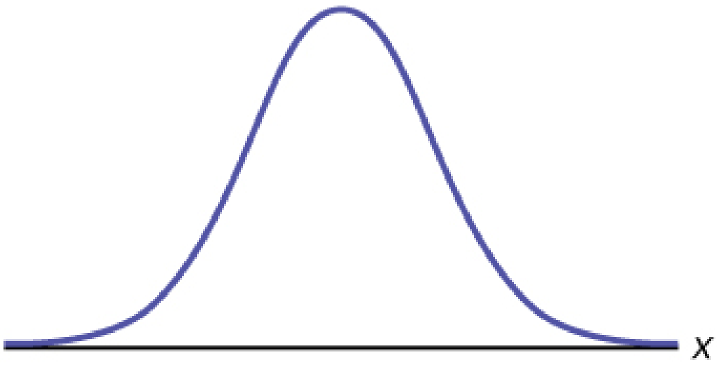
Problem: Complete the distributions.

a. X ~ _____(_____,_____)
b. X ~ _____(_____,_____)

Exercise:

Problem:

Find the probability that **one** review will take Yoonie from 3.5 to 4.25 hours. Sketch the graph, labeling and scaling the horizontal axis. Shade the region corresponding to the probability.



a.
b. $P($________ $< x <$ ________$) =$ _______

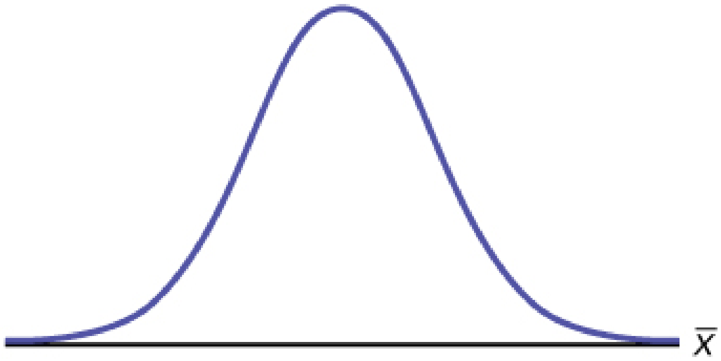
Solution:

a. Check student's solution.
b. 3.5, 4.25, 0.2441

Exercise:

Problem:

Find the probability that the **mean** of a month's reviews will take Yoonie from 3.5 to 4.25 hrs. Sketch the graph, labeling and scaling the horizontal axis. Shade the region corresponding to the probability.



a.

b. P(________________) = _______

Exercise:

Problem:

What causes the probabilities in [link] and [link] to be different?

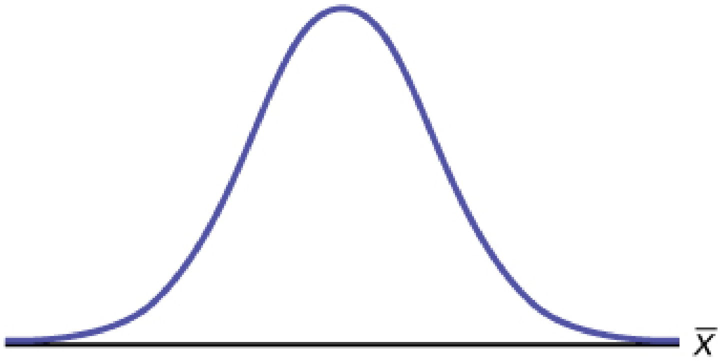
Solution:

The fact that the two distributions are different accounts for the different probabilities. In particular, the distributions have different standard deviations.

Exercise:

Problem:

Find the 95th percentile for the mean time to complete one month's reviews. Sketch the graph.



a.

b. The 95th Percentile =____________

Homework

Exercise:

Problem:

Previously, De Anza statistics students estimated that the amount of change daytime statistics students carry is normally distributed with a mean of $0.88 and a standard deviation of $0.25. Suppose that we randomly pick 35 daytime statistics students.

a. In words, X = _____________
b. X ~ _____(_____,_____)
c. In words, X = _____________
d. X ~ ______ (______, ______)
e. Find the probability that an individual had between $0.80 and $1.00. Graph the situation, and shade in the area to be determined.
f. Find the probability that the average of the 35 students was between $0.80 and $1.00. Graph the situation, and shade in the area to be determined.
g. Explain why there is a difference in part e and part f.

Solution:

a. X = amount of change students carry
b. X ~ $N(0.88, 0.25)$
c. X = average amount of change carried by a sample of 35 students.
d. X ~ $N(0.88, 0.0423)$
e. 0.3099
f. 0.9686
g. The distributions are different. While they are both normal and centered at 0.88, part a has a much larger standard deviation than the distribution in part b.

Exercise:

Problem:

Suppose that the distance of fly balls hit to the outfield (in baseball) is normally distributed with a mean of 250 feet and a standard deviation of 50 feet. We randomly sample 49 fly balls.

a. If X = average distance in feet for 49 fly balls, then X ~ ________(________,________)
b. What is the probability that the 49 balls traveled an average of less than 240 feet? Sketch the graph. Scale the horizontal axis for X. Shade the region corresponding to the probability. Find the probability.
c. Find the 80th percentile of the distribution of the average of 49 fly balls.

Exercise:

Problem:

According to the Internal Revenue Service, the average length of time for an individual to complete (keep records for, learn, prepare, copy, assemble, and send) IRS Form 1040 is 10.53 hours (without any attached schedules). The distribution is unknown. Let us assume that the standard deviation is two hours. Suppose we randomly sample 36 taxpayers.

a. In words, X = _______________
b. In words, X = _______________
c. X ~ ______(______,______)
d. Would you be surprised if the 36 taxpayers finished their Form 1040s in an average of more than 12 hours? Explain why or why not in complete sentences.
e. Would you be surprised if one taxpayer finished his or her Form 1040 in more than 12 hours? In a complete sentence, explain why.

Solution:

a. length of time for an individual to complete IRS form 1040, in hours.
b. mean length of time for a sample of 36 taxpayers to complete IRS form 1040, in hours.
c. $N\left(10.53,\ \frac{1}{3}\right)$
d. Yes. I would be surprised, because the probability is almost 0.
e. No. I would not be totally surprised because the probability is 0.2312

Exercise:

Problem:

Suppose that a category of world-class runners are known to run a marathon (26 miles) in an average of 145 minutes with a standard deviation of 14 minutes. Consider 49 of the races. Let X be the average of the 49 races.

a. X ~ _____(_____,_____)
b. Find the probability that the runner will average between 142 and 146 minutes in these 49 marathons.
c. Find the 80th percentile for the average of these 49 marathons.
d. Find the median of the average running times.

Exercise:

Problem:

The length of songs in a collector's iTunes album collection is uniformly distributed from two to 3.5 minutes. Suppose we randomly pick five albums from the collection. There are a total of 43 songs on the five albums.

a. In words, X = _________
b. X ~ _____________
c. In words, X = _____________

d. X ~ _____(_____,_____)
e. Find the first quartile for the average song length.
f. The IQR(interquartile range) for the average song length is from _______to_______.

Solution:

a. the length of a song, in minutes, in the collection
b. $U(2, 3.5)$
c. the average length, in minutes, of the songs from a sample of five albums from the collection
d. $N(2.75, 0.66)$
e. 2.71 minutes
f. 2.71 minutes to 2.79 minutes

Exercise:

Problem:

In 1940 the average size of a U.S. farm was 174 acres. Let’s say that the standard deviation was 55 acres. Suppose we randomly survey 38 farmers from 1940.

a. In words, X = ______________
b. In words, X = ______________
c. X ~ _____(_____,_____)
d. The IQR for X is from _______ acres to _______ acres.

Exercise:

Problem:

Determine which of the following are true and which are false. Then, in complete sentences, justify your answers.

a. When the sample size is large, the mean of X is equal to the mean of X.

b. When the sample size is large, X is approximately normally distributed.
c. When the sample size is large, the standard deviation of X is approximately the same as the standard deviation of X.

Solution:

a. True. The mean of a sampling distribution of the means is the same as the mean of the data distribution. Note that this is true regardless of the sample size.
b. True. According to the Central Limit Theorem, the larger the sample, the closer the sampling distribution of the means is to normal.
c. False. The standard deviation of the sampling distribution of the means will decrease making it smaller than the standard deviation of X as the sample size increases.

Exercise:

Problem:

The percent of fat calories that a person in America consumes each day is normally distributed with a mean of about 36 and a standard deviation of about ten. Suppose that 16 individuals are randomly chosen. Let X = average percent of fat calories.

a. X ~ ______(______, ______)
b. For the group of 16, find the probability that the average percent of fat calories consumed is more than five. Graph the situation and shade in the area to be determined.
c. Find the first quartile for the average percent of fat calories.

Exercise:

Problem:

The distribution of income in some Third World countries is considered wedge shaped (many very poor people, very few middle income people, and even fewer wealthy people). Suppose we pick a country with a wedge shaped distribution. Let the average salary be $2,000 per year with a standard deviation of $8,000. We randomly survey 1,000 residents of that country.

a. In words, X = _______________
b. In words, X = _______________
c. X ~ ______(______,______)
d. How is it possible for the standard deviation to be greater than the average?
e. Why is it more likely that the average of the 1,000 residents will be from $2,000 to $2,100 than from $2,100 to $2,200?

Solution:

a. X = the yearly income of someone in a third world country
b. the average salary from samples of 1,000 residents of a third world country
c. $X \sim N\left(2000, \frac{8000}{\sqrt{1000}}\right)$
d. Very wide differences in data values can have averages smaller than standard deviations.
e. The distribution of the sample mean will have higher probabilities closer to the population mean.
   $P(2000 < x < 2100)$ = 0.1537
   $P(2100 < x < 2200)$ = 0.1317

Exercise:

Problem:

Which of the following is NOT TRUE about the distribution for averages?

a. The mean, median, and mode are equal.
b. The area under the curve is one.
c. The curve never touches the x-axis.
d. The curve is skewed to the right.

Exercise:

Problem:

The cost of unleaded gasoline in the Bay Area once followed an unknown distribution with a mean of $4.59 and a standard deviation of $0.10. Fifty-six gas stations from the Bay Area are randomly chosen. We are interested in the average cost of gasoline for the 56 gas stations. The distribution to use for the average cost of gasoline for the 56 gas stations is:

a. $X \sim N(4.59, 0.10)$
b. $X \sim N\left(4.59, \frac{0.10}{\sqrt{56}}\right)$
c. $X \sim N\left(4.59, \frac{56}{0.10}\right)$
d. $X \sim N\left(4.59, \frac{\sqrt{56}}{0.10}\right)$

Solution:

b

Glossary

Average
: a number that describes the central tendency of the data; there are a number of specialized averages, including the arithmetic mean, weighted mean, median, mode, and geometric mean.

Sampling Distribution of the Sample Mean the distribution of every possible sample mean that can be obtained from selecting a random sample of size n from the original population (the distribution of X).

Central Limit Theorem

Given a random variable (RV) with known mean μ and known standard deviation, σ, we are sampling with size n, and we are interested in the new RV: the sample mean X. If the size (n) of the sample is sufficiently large (at least 30), then $X \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$. If the size ($n$) of the sample is sufficiently large, then the distribution of the sample means will approximate a normal distribution regardless of the shape of the population. The mean of the sample means will equal the population mean. The standard deviation of the distribution of the sample means, $\frac{\sigma}{\sqrt{n}}$, is called the standard error of the mean.

Standard Error of the Mean

the standard deviation of the distribution of the sample means, or $\frac{\sigma}{\sqrt{n}}$.

Using the Central Limit Theorem

Note:
NOTE
It is important for you to understand when to use the **central limit theorem**. If you are being asked to find the probability of a **mean**, then use the central limit theorem. This also applies to percentiles for means.
If you are being asked to find the probability of an **individual** value, do **not** use the central limit theorem. **Use the distribution of its random variable.**

Examples of the Central Limit Theorem

Law of Large Numbers

The **law of large numbers** says that if you take samples of larger and larger size from any population, then the mean x of the sample tends to get closer and closer to μ. From the central limit theorem, we know that as n gets larger and larger, the sample means follow a normal distribution. The larger n gets, the smaller the standard deviation gets. (Remember that the standard deviation for X is $\frac{\sigma}{\sqrt{n}}$.) This means that the sample mean x must be close to the population mean μ. We can say that μ is the value that the sample means approach (cluster more tightly around) as n gets larger. The central limit theorem illustrates the law of large numbers.

Central Limit Theorem for the Mean Examples

Example:
A study involving stress is conducted among the students on a college campus. **The stress scores follow a uniform distribution** with the lowest stress score equal to one and the highest equal to five. Using a sample of 75 students, find:

a. The probability that the **mean stress score** for the 75 students is less than two.
b. The 90th percentile for the **mean stress score** for the 75 students.

Let X = one stress score.
Parts a and b ask you to find a probability or a percentile for a **mean**.
Since the individual stress scores follow a uniform distribution, $X \sim U(1, 5)$ where $a = 1$ and $b = 5$ (See The Uniform Distribution for an explanation on the uniform distribution).
$\mu_x = \frac{a+b}{2} = \frac{1+5}{2} = 3$
$\sigma_x = \sqrt{\frac{(b–a)^2}{12}} = \sqrt{\frac{(5–1)^2}{12}} = 1.15$
Let X = the mean stress score for the 75 students. Then,
$X \sim N\left(3, \frac{1.15}{\sqrt{75}}\right)$
Exercise:

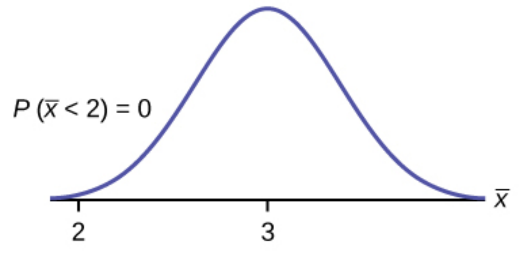
Problem: a. Find $P(x < 2)$. Draw the graph.

Solution:

a. $P(x < 2) = 0$

`normalcdf` $\left(1,2,3,\frac{1.15}{\sqrt{75}}\right)$ = 0

The probability that the mean stress score is less than two is about zero.



Note:
Reminder
The smallest stress score is one.

Exercise:

Problem: b. Find the 90th percentile for the mean of 75 stress scores. Draw a graph.

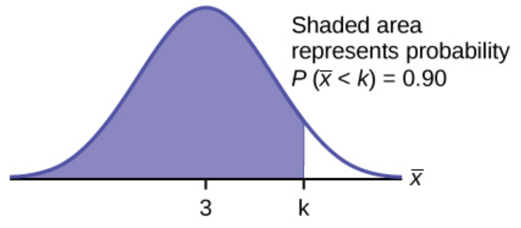
Solution:

b. Let k = the 90th precentile.

Find k, where $P(x < k)$ = 0.90.

k = 3.2

`invNorm` $\left(0.90,3,\frac{1.15}{\sqrt{75}}\right)$ = 3.2



The 90th percentile for the mean of 75 scores is about 3.2. This tells us that 90% of all the means of 75 stress scores are at most 3.2, and that 10% are at least 3.2.

Note:
Try It
Exercise:

Problem: Use the information in [link], but use a sample size of 55 to answer the following questions.

a. Find $P(x < 2.7)$.
b. Find the 80th percentile for the mean of 55 scores.

Solution:
Solutions

a. 0.0265
b. 3.13

Example:
In the United States, someone is sexually assaulted every two minutes, on average, according to a number of studies. Suppose the standard deviation is 0.5 minutes and the sample size is 100.
Exercise:

Problem:

a. Find the median, the first quartile, and the third quartile for the sample mean time of sexual assaults in the United States.
b. Find the probability that a sexual assault occurs on the average between 1.75 and 1.85 minutes.
c. Find the value that is two standard deviations above the sample mean.

Solution:

a. We have, $\mu_x = \mu_x = 2$ and $\sigma_x = \frac{\sigma_x}{\sqrt{n}} = \frac{0.5}{10} = 0.05$. Therefore:

    1. median = $\mu_x = 2$
    2. 25th percentile = *invNorm*(0.25,2,0.05) = 1.97
    3. 75th percentile = *invNorm*(0.75,2,0.05) = 2.03

b. $P(1.75 < x < 1.85)$ = `normalcdf`(1.75,1.85,2,0.05) = 0.0013
c. Using the z-score formula, $z_x = \frac{x-\mu_x}{\sigma_x}$, and solving for x, we have $x = 2(0.05) + 2 = 2.1$

Note:
Try It
Exercise:

Problem:

Based on data from the National Health Survey, women between the ages of 18 and 24 have an average systolic blood pressures (in mm Hg) of 114.8 with a standard deviation of 13.1. Systolic blood pressure for women between the ages of 18 to 24 follow a normal distribution.

a. If one woman from this population is randomly selected, find the probability that her systolic blood pressure is greater than 120.
b. If 40 women from this population are randomly selected, find the probability that their mean systolic blood pressure is greater than 120.

c. If the sample were four women between the ages of 18 to 24 and we did not know the original distribution, could the central limit theorem be used?

Solution:

a. $P(x > 120)$ = `normalcdf`(120,1E99,114.8,13.1) = 0.3457. There is about a 35%, that the randomly selected woman will have systolics blood pressure greater than 120.
b. $P(x > 120)$ = `normalcdf`$\left(120,1E99,114.8,\frac{13.1}{\sqrt{40}}\right)$ = 0.006. There is only a 0.6% chance that the average systolic blood pressure for the randomly selected group is greater than 120.
c. The central limit theorem could not be used if the sample size were four and we did not know the original distribution was normal. The sample size would be too small.

Example:
Exercise:

Problem:

A study was done about violence against prostitutes and the symptoms of the post traumatic stress that they developed. The age range of the prostitutes was 14 to 61. The mean age was 30.9 years with a standard deviation of nine years.

a. In a sample of 25 prostitutes, what is the probability that the mean age of the prostitutes is less than 35?
b. Is it likely that the mean age of the sample group could be more than 50 years? Interpret the results.
c. Find the 95^{th} percentile for the sample mean age of 65 prostitutes. Interpret the results.

Solution:

a. $P(x < 35)$ = `normalcdf`(-1E99,35,30.9,1.8) = 0.9886
b. $P(x > 50)$ = `normalcdf`(50, 1E99,30.9,1.8) ≈ 0. For this sample group, it is almost impossible for the group's average age to be more than 50. However, it is still possible for an individual in this group to have an age greater than 50.
c. The 95th percentile = `invNorm`(0.95,30.9,1.1) = 32.7. This indicates that 95% of samples of 65 prostitutes will have an average age younger than 32.7 years.

Note:
Try It
Exercise:

Problem:

According to Boeing data, the 757 airliner carries 200 passengers and has doors with a height of 72 inches. Assume for a certain population of men we have a mean height of 69.0 inches and a standard deviation of 2.8 inches.

a. What doorway height would allow 95% of men to enter the aircraft without bending?

b. Assume that half of the 200 passengers are men. What mean doorway height satisfies the condition that there is a 0.95 probability that this height is greater than the mean height of 100 men?
c. For engineers designing the 757, which result is more relevant: the height from part a or part b? Why?

Solution:

a. We know that μ_x = 69 and we have σ_x = 2.8. The height of the doorway is found to be `invNorm`(0.95,69,2.8) = 73.61
b. We know that $\mu_x = \mu_x$ = 69 and we have $\sigma_x = \frac{\sigma_x}{\sqrt{n}}$ = $\frac{2.8}{10}$ = 0.28. So, `invNorm`(0.95,69,0.28) = 69.46
c. When designing the doorway heights, we need to incorporate as much variability as possible in order to accommodate as many passengers as possible. Therefore, we need to use the result based on part a.

References

Data from the Wall Street Journal.

"National Health and Nutrition Examination Survey." Center for Disease Control and Prevention. Available online at http://www.cdc.gov/nchs/nhanes.htm (accessed May 17, 2013).

Chapter Review

The central limit theorem can be used to illustrate the law of large numbers. The law of large numbers states that the larger the sample size you take from a population, the closer a single sample mean x gets to μ.

Use the following information to answer the next six exercises: A manufacturer produces 25-pound lifting weights. The lowest actual weight is 24 pounds, and the highest is 26 pounds. Each weight is equally likely so the distribution of weights is uniform. A sample of 100 weights is taken.

Exercise:

Problem:

a. What is the distribution for the weights of one 25-pound lifting weight? What is the mean and standard deivation?
b. What is the distribution for the mean weight of 100 25-pound lifting weights?
c. Find the probability that the mean actual weight for the 100 weights is less than 24.9.

Solution:

a. $U(24, 26)$, 25, 0.577
b. $N(25, 0.0577)$
c. 0.0415

Exercise:

Problem: Draw the graph from [link]

Exercise:

Problem: Find the probability that the mean actual weight for the 100 weights is greater than 25.2.

Solution:

0.0003

Exercise:

Problem: Draw the graph from [link]

Exercise:

Problem: Find the 90th percentile for the mean weight for the 100 weights.

Solution:

25.07

Exercise:

Problem: Draw the graph from [link]

Homework

Use the following information to answer the next seven exercises: Richard’s Furniture Company delivers furniture from 10 A.M. to 2 P.M. continuously and uniformly. We are interested in how long (in hours) past the 10 A.M. start time that individuals wait for their delivery.

Exercise:

Problem: X ~ ______(______,______)

a. $U(0,4)$
b. $U(10,2)$
c. $N(2,1.15)$

Exercise:

Problem:

Suppose a random sample of 50 customers are selected and X represents the average wait time of the sample.

Then, X ~ ______(______,______)

a. $U(0,4)$
b. $N(2,1.15)$
c. $N(2,0.163)$

Solution:

c

Exercise:

Problem: The average wait time for an individual customer is:

a. one hour.
b. two hours.
c. two and a half hours.
d. four hours.

Exercise:

Problem: The probability a customer will wait more than 3.5 hours is

a. $\frac{1}{8}$
b. $\frac{1}{2}$
c. $\frac{3}{4}$
d. $\frac{3}{8}$

Solution:

a

Exercise:

Problem:

Suppose that it is now past noon on a delivery day. The probability that a person must wait at least one and a half **more** hours is:

a. $\frac{1}{8}$
b. $\frac{1}{2}$
c. $\frac{1}{4}$
d. $\frac{3}{8}$

Exercise:

Problem:

What's the probability that the average wait time for 50 randomly selected customers is more than 3.5 hours?

Solution:

0

Exercise:

Problem: Write out in complete sentences why the answers to the last three exercises are different.

Use the following information to answer the next two exercises: The time to wait for a particular rural bus is distributed uniformly from zero to 75 minutes. One hundred riders are randomly sampled to learn how long they waited.

Exercise:

Problem: The 90th percentile sample average wait time (in minutes) for a sample of 100 riders is:

a. 315.0
b. 40.3
c. 38.5
d. 65.2

Solution:

b

Exercise:

Problem:

Would you be surprised, based upon numerical calculations, if the sample average wait time (in minutes) for 100 riders was less than 30 minutes?

a. yes
b. no
c. There is not enough information.

Use the following to answer the next two exercises: The cost of unleaded gasoline in the Bay Area once followed an unknown distribution with a mean of \$4.59 and a standard deviation of \$0.10. Fifty-six gas stations from the Bay Area are randomly chosen. We are interested in the average cost of gasoline for the 56 gas stations.

Exercise:

Problem: What's the approximate probability that the average price for 56 gas stations is over \$4.69?

a. almost zero
b. 0.1587
c. 0.0943
d. unknown

Solution:

a

Exercise:

Problem: Find the probability that the average price for 56 gas stations is less than \$4.55.

a. 0.6554
b. 0.3446
c. 0.0014
d. 0.9858
e. 0

Exercise:

Problem:

$X \sim N(60, 9)$. Suppose that you form random samples of 25 from this distribution. Let X be the random variable of averages. For parts c through f, sketch the graph, shade the region, label and scale the horizontal axis for X, and find the probability.

a. Sketch the distributions of X and X.
b. $X \sim$ ______(______,______)
c. $P(x < 60)$ = ______
d. Find the 30th percentile for the mean.
e. $P(56 < x < 62)$ = ______
f. $P(18 < x < 58)$ = ______

Solution:

a. Check student's solution.
b. $X \sim N\left(60, \frac{9}{\sqrt{25}}\right)$
c. 0.5000
d. 59.06
e. 0.8536
f. 0.1333

Exercise:

Problem:

Suppose that the length of research papers is uniformly distributed from ten to 25 pages. We survey a class in which 55 research papers were turned in to a professor. The 55 research papers are considered a random collection of all papers. We are interested in the average length of the research papers.

a. In words, X = ______________
b. $X \sim$ ______(______,______)
c. μ_x = ______
d. σ_x = ______
e. In words, X = _______________
f. $X \sim$ ______(______,______)

Exercise:

Problem:

Salaries for teachers in a particular elementary school district are normally distributed with a mean of $44,000 and a standard deviation of $6,500. We randomly survey ten teachers from that district.

a. Find the 90th percentile for an individual teacher's salary.
b. Find the 90th percentile for the average teacher's salary.

Solution:

a. $52,330
b. $46,634

Exercise:

Problem:

The average length of a maternity stay in a U.S. hospital is said to be 2.4 days with a standard deviation of 0.9 days. We randomly survey 80 women who recently bore children in a U.S. hospital.

a. In words, X = ______________
b. In words, X = ____________________
c. X ~ _____(_____,_____)
d. Is it likely that an individual stayed more than five days in the hospital? Why or why not?
e. Is it likely that the average stay for the 80 women was more than five days? Why or why not?
f. Which is more likely:

    i. An individual stayed more than five days.
    ii. the average stay of 80 women was more than five days.

For each problem, wherever possible, provide graphs and use the calculator.

Exercise:

Problem:

NeverReady batteries has engineered a newer, longer lasting AAA battery. The company claims this battery has an average life span of 17 hours with a standard deviation of 0.8 hours. Your statistics class questions this claim. As a class, you randomly select 30 batteries and find that the sample mean life span is 16.7 hours. If the process is working properly, what is the probability of getting a random sample of 30 batteries in which the sample mean lifetime is 16.7 hours or less? Is the company's claim reasonable?

Solution:

- We have μ = 17, σ = 0.8, x = 16.7, and n = 30. To calculate the probability, we use `normalcdf`(lower, upper, μ, $\frac{\sigma}{\sqrt{n}}$) = `normalcdf`$\left(-1\text{ E}99, 16.7, 17, \frac{0.8}{\sqrt{30}}\right)$ = 0.0200.
- If the process is working properly, then the probability that a sample of 30 batteries would have at most 16.7 lifetime hours is only 2%. Therefore, the class was justified to question the claim.

Exercise:

Problem:

The Screw Right Company claims their $\frac{3}{4}$ inch screws are within ±0.23 of the claimed mean diameter of 0.750 inches with a standard deviation of 0.115 inches. The following data were recorded.

0.757	0.723	0.754	0.737	0.757	0.741	0.722	0.741	0.743	0.742
0.740	0.758	0.724	0.739	0.736	0.735	0.760	0.750	0.759	0.754
0.744	0.758	0.765	0.756	0.738	0.742	0.758	0.757	0.724	0.757

0.744	0.738	0.763	0.756	0.760	0.768	0.761	0.742	0.734	0.754
0.758	0.735	0.740	0.743	0.737	0.737	0.725	0.761	0.758	0.756

The screws were randomly selected from the local home repair store.

a. Find the mean diameter and standard deviation for the sample
b. Find the probability that 50 randomly selected screws will be within the stated tolerance levels. Is the company's diameter claim plausible?

Exercise:

Problem:

Your company has a contract to perform preventive maintenance on thousands of air-conditioners in a large city. Based on service records from previous years, the time that a technician spends servicing a unit averages one hour with a standard deviation of one hour. In the coming week, your company will service a simple random sample of 70 units in the city. You plan to budget an average of 1.1 hours per technician to complete the work. Will this be enough time?

Solution:

Use `normalcdf`$\left(-1\text{ E}99,1.1,1,\frac{1}{\sqrt{70}}\right) = 0.7986$. This means that there is an 80% chance that the service time will be less than 1.1 hours. It could be wise to schedule more time since there is an associated 20% chance that the maintenance time will be greater than 1.1 hours.

Exercise:

Problem:

A typical adult has an average IQ score of 105 with a standard deviation of 20. If 20 randomly selected adults are given an IQ test, what is the probability that the sample mean scores will be between 85 and 125 points?

Confidence Intervals: Introduction class="introduction"

Have you ever wondered what the average number of M&Ms in a bag at the grocery store is? You can use confidence intervals to answer this question. (credit: comedy_nose/flickr)

Note:
Chapter Objectives
By the end of this chapter, the student should be able to:

- Calculate and interpret confidence intervals for estimating a population mean and a population proportion.
- Interpret the Student's t probability distribution as the sample size changes.
- Discriminate between problems applying the Normal and the Student's t distributions.
- Calculate the sample size required to estimate a population mean and a population proportion given a desired confidence level and margin of error.

Suppose you were trying to determine the mean rent of a two-bedroom apartment in your town. You might look in the classified section of the newspaper, write down several rents listed, and average them together. You would have obtained a point estimate of the true mean. If you are trying to determine the percentage of times you make a basket when shooting a basketball, you might count the number of shots you make and divide that by the number of shots you attempted. In this case, you would have obtained a point estimate for the true proportion.

We use sample data to make generalizations about an unknown population. This part of statistics is called **inferential statistics**. **The sample data help us to make an estimate of a population parameter**. We realize that the point estimate is most likely not the exact value of the population parameter, but close to it. After calculating point estimates, we construct interval estimates, called confidence intervals.

In this chapter, you will learn to construct and interpret confidence intervals. You will also learn a new distribution, the Student's t-distribution, and how it is used with these intervals. Throughout the chapter, it is important to keep in mind that the confidence interval is a random variable. It is the population parameter that is fixed.

If you worked in the marketing department of an entertainment company, you might be interested in the mean number of songs a consumer downloads a month from iTunes. If so, you could conduct a survey and calculate the sample mean, x, and the sample standard deviation, s. You would use x to estimate the population mean and s to estimate the population standard deviation. The sample mean, x, is the **point estimate** for the population mean, μ. The sample standard deviation, s, is the point estimate for the population standard deviation, σ.

Each of x and s is called a statistic.

A **confidence interval** is another type of estimate but, instead of being just one number, it is an interval of numbers. It provides a range of reasonable values in which we expect the population parameter to fall. There is no guarantee that a given confidence interval does capture the parameter, but there is a predictable probability of success.

Suppose, for the iTunes example, we do not know the population mean μ, but we do know that the population standard deviation is $\sigma = 1$ and our sample size is 100. Then, by the central limit theorem, the standard deviation for the sample mean is

$$\frac{\sigma}{\sqrt{n}} = \frac{1}{\sqrt{100}} = 0.1.$$

The **empirical rule**, which applies to bell-shaped distributions, says that in approximately 95% of the samples, the sample mean, x, will be within two standard deviations of the population mean μ. For our iTunes example, two standard deviations is (2)(0.1) = 0.2. The sample mean x is likely to be within 0.2 units of μ.

Because x is within 0.2 units of μ, which is unknown, then μ is likely to be within 0.2 units of x in 95% of the samples. The population mean μ is contained in an interval whose lower number is calculated by taking the sample mean and subtracting two standard deviations (2)(0.1) and whose upper number is calculated by taking the sample mean and adding two standard deviations. In other words, μ is between $x\ -\ 0.2$ and $x\ +\ 0.2$ in 95% of all the samples.

For the iTunes example, suppose that a sample produced a sample mean $x\ =\ 2$. Then the unknown population mean μ is between

$x - 0.2 = 2 - 0.2 = 1.8$ and $x + 0.2 = 2 + 0.2 = 2.2$

We say that we are **95% confident** that the unknown population mean number of songs downloaded from iTunes per month is between 1.8 and 2.2. **The 95% confidence interval is (1.8, 2.2).**

The 95% confidence interval implies two possibilities. Either the interval (1.8, 2.2) contains the true mean μ or our sample produced an x that is not within 0.2 units of the true mean μ. The second possibility happens for only 5% of all the samples (100-95%).

Remember that a confidence interval is created for an unknown population parameter like the population mean, μ. Confidence intervals for some parameters have the form:

(point estimate – margin of error, point estimate + margin of error)

The margin of error depends on the confidence level or percentage of confidence and the standard error of the mean.

When you read newspapers and journals, some reports will use the phrase "margin of error." Other reports will not use that phrase, but include a confidence interval as the point estimate plus or minus the margin of error. These are two ways of expressing the same concept.

Note:
Note
Although the text only covers symmetrical confidence intervals, there are non-symmetrical confidence intervals (for example, a confidence interval for the standard deviation).

Note:
Collaborative Exercise
Have your instructor record the number of meals each student in your class eats out in a week. Assume that the standard deviation is known to be three meals. Construct an approximate 95% confidence interval for the true mean number of meals students eat out each week.

1. Calculate the sample mean.
2. Let σ = 3 and n = the number of students surveyed.
3. Construct the interval $\left(x - 2\cdot\frac{\sigma}{\sqrt{n}}, x + 2\cdot\frac{\sigma}{\sqrt{n}}\right)$.

We say we are approximately 95% confident that the true mean number of meals that students eat out in a week is between ___________ and ____________.

Glossary

Confidence Interval (CI)

an interval estimate for an unknown population parameter. This depends on:

- the desired confidence level,
- information that is known about the distribution (for example, known standard deviation),
- the sample and its size.

Inferential Statistics

also called statistical inference or inductive statistics; this facet of statistics deals with estimating a population parameter based on a sample statistic. For example, if four out of the 100 calculators sampled are defective we might infer that four percent of the production is defective.

Parameter

a numerical characteristic of a population

Point Estimate

a single number computed from a sample and used to estimate a population parameter

Estimating a Single Population Mean using the Normal Distribution

A confidence interval for a population mean with a known standard deviation is based on the fact that the sample means follow an approximately normal distribution. Suppose that our sample has a mean of x = 10 and we have constructed the 90% confidence interval (5, 15) where *EBM* = 5.

Calculating the Confidence Interval

To construct a confidence interval for a single unknown population mean μ, **where the population standard deviation is known**, we need x as an estimate for μ and we need the margin of error. Here, the margin of error (*EBM*) is called the **error bound for a population mean** (abbreviated ***EBM***). The sample mean x is the **point estimate** of the unknown population mean μ.

The confidence interval estimate will have the form:

(lower bound, upper bound) or (point estimate - error bound, point estimate + error bound), or, in symbols,$(x-EBM, x+EBM)$

The margin of error (*EBM*) depends on the **confidence level** (abbreviated CL). The confidence level is the percent of confidence intervals, constructed using the same method, that will contain the true population parameter, when many repeated samples are taken. Thus, the confident level represents how confident we can be that a particular confident interval has captured the true population parameter being estimated. Most often, it is the choice of the person constructing the confidence interval to choose a confidence level of 90% or higher because that person wants to be reasonably certain of his or her conclusions.

There is another percentage called alpha (α). α is related to the confidence level, CL. α is the percentage of confidence intervals that will not contain the unknown population parameter, in repeated sampling.
Mathematically, α = 1 - CL.

Example:

- Suppose we have collected data from a sample. We know the sample mean, but we do not know the mean for the entire population.
- The sample mean is seven, and the error bound for the mean is 2.5.

x = 7 and *EBM* = 2.5
The confidence interval is (7 – 2.5, 7 + 2.5), and simplifying this gives us (4.5, 9.5).
If the confidence level (CL) is 95%, then we say, "We estimate with 95% confidence that the true value of the population mean is between 4.5 and 9.5."

Note:
Try It
Exercise:

Problem:

Suppose we have data from a sample. The sample mean is 15, and the error bound for the mean is 3.2.

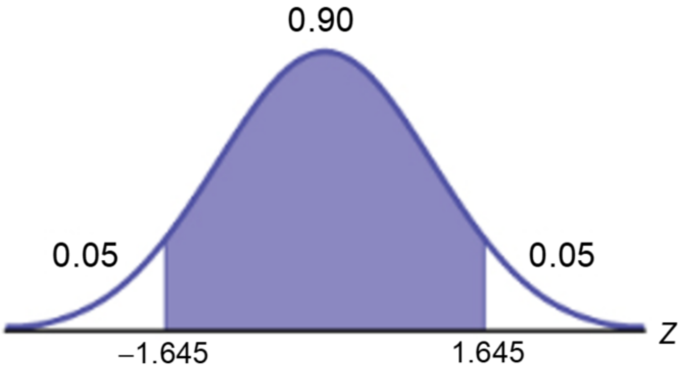
What is the confidence interval estimate for the population mean?

Solution:

(11.8, 18.2)

A confidence interval for a population mean with a known standard deviation is based on the fact that the sample means follow an approximately normal distribution. Suppose that our sample has a mean of x = 10, and we have constructed the 90% confidence interval (5, 15) where *EBM* = 5.

The EBM is calculated using a formula that is based on the z-score associated with the confidence level. To find the z-score for a 90% confidence interval, we must consider the central 90% of the probability of the standard normal distribution. If we include the central 90%, we leave out a total of α = 10% in both tails, or 5% in each tail, of the standard normal distribution.



The value 1.645 is the z-score from a standard normal probability distribution that puts an area of 0.90 in the center, an area of 0.05 in the far left tail, and an area of 0.05 in the far right tail. This z-score is then multiplied to the appropriate standard deviation to calculate the error bound.

It is important that the "standard deviation" used be appropriate for the parameter we are estimating, so in this section we need to use the standard deviation that applies to sample means, which is $\frac{\sigma}{\sqrt{n}}$. The fraction $\frac{\sigma}{\sqrt{n}}$, is commonly called the "standard error of the mean", in order to distinguish clearly the standard deviation of the sample means from the population standard deviation σ.

By the central limit theorem:

- X is normally distributed, and in particular, $X \sim N\left(\mu_x, \frac{\sigma}{\sqrt{n}}\right)$.
- **Therefore, when the population standard deviation σ is known, we use a normal distribution to calculate the error bound.**

Calculating the Confidence Interval

To construct a confidence interval estimate for an unknown population mean, we need data from a random sample. The steps to construct and interpret the confidence interval are:

- Calculate the sample mean x from the sample data. Remember, in this section we already know the population standard deviation σ.
- Find the z-score that corresponds to the confidence level.

- Calculate the error bound, *EBM*, by multiplying the z-score by the standard error, $\frac{\sigma}{\sqrt{n}}$.
- Construct the confidence interval, $(x - EBM, x + EBM)$.
- Write a sentence that interprets the estimate in the context of the situation in the problem. (Explain what the confidence interval means, in the words of the problem.)

We will first examine each step in more detail, and then illustrate the process with some examples.

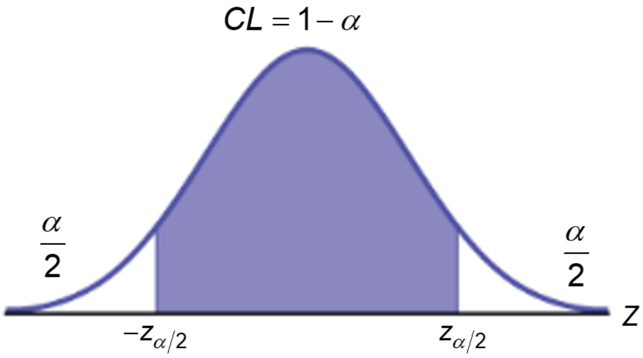
Finding the z-score for the Stated Confidence Level

When we know the population standard deviation σ, we use a standard normal distribution to calculate the error bound EBM and construct the confidence interval. We need to find the value of z that puts an area equal to the confidence level (in decimal form) in the middle of the standard normal distribution Z ~ $N(0, 1)$.

The confidence level, CL, is the area in the middle of the standard normal distribution. CL = 1 – α, so α is the area that is split equally between the two tails. Each of the tails contains an area equal to $\frac{\alpha}{2}$.

The z-score that has an area to the right of $\frac{\alpha}{2}$ is denoted by $z_{\frac{\alpha}{2}}$.

For example, when CL = 0.95, α = 0.05 and $\frac{\alpha}{2}$ = 0.025; we write $z_{\frac{\alpha}{2}}$ = $z_{0.025}$.



The area to the right of $z_{0.025}$ is 0.025 and the area to the left of $z_{0.025}$ is 1 – 0.025 = 0.975.

$z_{\frac{\alpha}{2}} = z_{0.025} = 1.96$, using a calculator, computer, or a standard normal probability table.

Note:
`invNorm`(0.975, 0, 1) = 1.96

Note:
Note
Remember to use the area to the LEFT of $z_{\frac{\alpha}{2}}$; in this chapter the last two inputs in the invNorm command are 0, 1, because you are using a standard normal distribution $Z \sim N(0, 1)$.

Calculating the Error Bound (*EBM*)

The error bound formula for an unknown population mean μ when the population standard deviation σ is known is

- $EBM = \left(z_{\frac{\alpha}{2}}\right)\left(\frac{\sigma}{\sqrt{n}}\right)$

Constructing the Confidence Interval

- The confidence interval has the format $(x - EBM, x + EBM)$.

Writing the Interpretation

The interpretation should clearly state the confidence level (CL), explain what population parameter is being estimated (here, a **population mean**), and state the confidence interval (both endpoints). "We estimate with ___% confidence that the true population mean (include the context of the problem) is between ___ and ___ (include appropriate units)."

Example:
Suppose scores on exams in statistics are normally distributed with an unknown population mean and a population standard deviation of three points. A random sample of 36 scores is taken and gives a sample mean (sample mean score) of 68. Find a confidence interval to estimate the population mean exam score (the mean score on all exams).
Exercise:

Problem:

Find a 90% confidence interval for the true (population) mean of statistics exam scores.

- You can use technology to calculate the confidence interval directly.
- The first solution is shown step-by-step (Solution A).
- The second solution uses the TI-83, 83+, and 84+ calculators (Solution B).

Solution:

To find the confidence interval, you need the sample mean, x, and the *EBM*.

- x = 68

- $EBM = \left(z_{\frac{\alpha}{2}}\right)\left(\frac{\sigma}{\sqrt{n}}\right)$
- σ = 3; n = 36; the confidence level is 90% (CL = 0.90)

CL = 0.90 so α = 1 – CL = 1 – 0.90 = 0.10

$\frac{\alpha}{2}$ = 0.05; $z_{\frac{\alpha}{2}} = z_{0.05}$

The area to the right of $z_{0.05}$ is 0.05 and the area to the left of $z_{0.05}$ is 1 – 0.05 = 0.95.

$z_{\frac{\alpha}{2}} = z_{0.05} = 1.645$

Note that we find the z-score 1.645 using invNorm(0.95, 0, 1) on the TI-83, 83+, and 84+ calculators. This can also be found using appropriate commands on other calculators, using a computer, or using a probability table for the standard normal distribution.

$EBM = (1.645)\left(\frac{3}{\sqrt{36}}\right) = 0.8225$

x - *EBM* = 68 - 0.8225 = 67.1775

x + *EBM* = 68 + 0.8225 = 68.8225

The 90% confidence interval is **(67.1775, 68.8225).**

Solution:

Note:
Press `STAT` and arrow over to `TESTS`.
Arrow down to `7:ZInterval`.
Press `ENTER`.
Arrow to `Stats` and press `ENTER`.
Arrow down and enter 3 for σ, 68 for x, 36 for n, and .90 for `C-level`.
Arrow down to `Calculate` and press `ENTER`.
The confidence interval is (67.178, 68.822).

Interpretation
We estimate with 90% confidence that the true mean exam score for all statistics students is between 67.18 and 68.82.

Explanation of 90% Confidence Level (Why can we be 90% confident?)
Ninety percent of all confidence intervals constructed in this way contain the true mean statistics exam score. For example, if we constructed 100 of these confidence intervals, we would expect 90 of them to contain the true population mean exam score.

Note:
Try It
Suppose average pizza delivery times are normally distributed with an unknown population mean and a known population standard deviation of six minutes. A random sample of 28 pizza delivery restaurants is taken and has a sample mean delivery time of 36 minutes.
Exercise:

Problem:

Find a 90% confidence interval estimate for the population mean delivery time.

Solution:

(34.1347, 37.8653)

Example:
The Specific Absorption Rate (SAR) for a cell phone measures the amount of radio frequency (RF) energy absorbed by the user’s body when using the handset. Every cell phone emits RF energy. Different phone models have different SAR measures. To receive certification from the Federal Communications Commission (FCC) for sale in the United States, the SAR level for a cell phone must be no more than 1.6 watts per kilogram. The following table shows the highest SAR level for a random selection of cell

phone models as measured by the FCC. Assume SAR measurements are normally distributed.

Phone Model	**SAR**	**Phone Model**	**SAR**	**Phone Model**	**SAR**
Apple iPhone 4S	1.11	LG Ally	1.36	Pantech Laser	0.74
BlackBerry Pearl 8120	1.48	LG AX275	1.34	Samsung Character	0.5
BlackBerry Tour 9630	1.43	LG Cosmos	1.18	Samsung Epic 4G Touch	0.4
Cricket TXTM8	1.3	LG CU515	1.3	Samsung M240	0.867
HP/Palm Centro	1.09	LG Trax CU575	1.26	Samsung Messager III SCH-R750	0.68
HTC One V	0.455	Motorola Q9h	1.29	Samsung Nexus S	0.51
HTC Touch Pro 2	1.41	Motorola Razr2 V8	0.36	Samsung SGH-A227	1.13
Huawei M835 Ideos	0.82	Motorola Razr2 V9	0.52	SGH-a107 GoPhone	0.3

Phone Model	SAR	Phone Model	SAR	Phone Model	SAR
Kyocera DuraPlus	0.78	Motorola V195s	1.6	Sony W350a	1.48
Kyocera K127 Marbl	1.25	Nokia 1680	1.39	T-Mobile Concord	1.38

Exercise:

Problem:

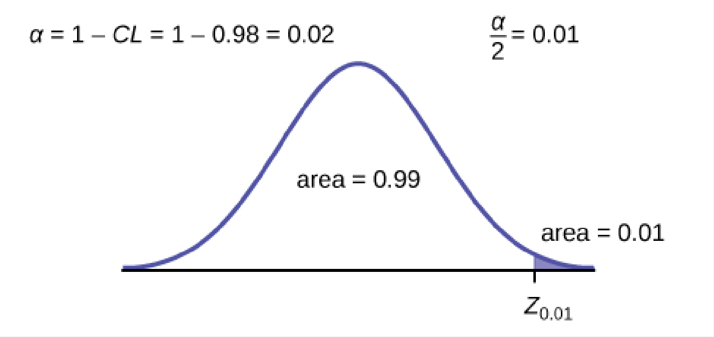
Find a 98% confidence interval for the true (population) mean of the Specific Absorption Rates (SARs) for cell phones. Assume that the population standard deviation is σ = 0.337.

Solution:

To find the confidence interval, start by finding the point estimate: the sample mean.

$x = 1.024$

Next, find the *EBM*. Because you are creating a 98% confidence interval, CL = 0.98.


$\alpha = 1 – CL = 1 – 0.98 = 0.02$
$\frac{\alpha}{2} = 0.01$
area = 0.99
area = 0.01
$z_{0.01}$

You need to find $z_{0.01}$ having the property that the area under the normal density curve to the right of $z_{0.01}$ is 0.01 and the area to the left is 0.99. Use your calculator, a computer, or a probability table for the standard normal distribution to find $z_{0.01}$ = 2.326.

$$EBM = (z_{0.01})\frac{\sigma}{\sqrt{n}} = (2.326)\frac{0.337}{\sqrt{30}} = 0.1431$$

To find the 98% confidence interval, find $x \pm EBM$.

x – *EBM* = 1.024 – 0.1431 = 0.8809

x + *EBM* = 1.024 + 0.1431 = 1.1671

We estimate with 98% confidence that the true SAR mean for the population of cell phones in the United States is between 0.8809 and 1.1671 watts per kilogram.

Solution:

Note:

- Press STAT and arrow over to TESTS.
- Arrow down to 7:ZInterval.
- Press ENTER.
- Arrow to Stats and press ENTER.
- Arrow down and enter the following values:
  - σ: 0.337
  - x : 1.024
  - n: 30
  - C-level: 0.98
- Arrow down to Calculate and press ENTER.
- The confidence interval is (0.88087, 1.1671).

Note:
Try It

Exercise:

Problem:

The following table shows a different random sampling of 20 cell phone models. Use this data to calculate a 93% confidence interval for the true mean SAR for cell phones certified for use in the United States. As previously, assume that SAR measurements are normally distributed and the population standard deviation is $\sigma = 0.337$.

Phone Model	**SAR**	**Phone Model**	**SAR**
Blackberry Pearl 8120	1.48	Nokia E71x	1.53
HTC Evo Design 4G	0.8	Nokia N75	0.68
HTC Freestyle	1.15	Nokia N79	1.4
LG Ally	1.36	Sagem Puma	1.24
LG Fathom	0.77	Samsung Fascinate	0.57
LG Optimus Vu	0.462	Samsung Infuse 4G	0.2
Motorola Cliq XT	1.36	Samsung Nexus S	0.51
Motorola Droid Pro	1.39	Samsung Replenish	0.3
Motorola Droid Razr M	1.3	Sony W518a Walkman	0.73
Nokia 7705 Twist	0.7	ZTE C79	0.869

Solution:

$x = 0.940$

$\frac{\alpha}{2} = \frac{1-CL}{2} = \frac{1-0.93}{2} = 0.035$

$z_{0.035} = 1.812$

$$EBM = (z_{0.035})\left(\frac{\sigma}{\sqrt{n}}\right) = (1.812)\left(\frac{0.337}{\sqrt{20}}\right) = 0.1365$$

$x – EBM = 0.940 – 0.1365 = 0.8035$

$x + EBM = 0.940 + 0.1365 = 1.0765$

We estimate with 93% confidence that the true SAR mean for the population of cell phones in the United States is between 0.8035 and 1.0765 watts per kilogram.

Notice the difference in the confidence intervals calculated in [link] and the following Try It exercise. These intervals are different for several reasons: they were calculated from different samples, the samples were different sizes, and the intervals were calculated for different levels of confidence. Even though the intervals are different, they do not yield conflicting information. The effects of these kinds of changes are the subject of the next section in this chapter.

Changing the Confidence Level or Sample Size

Example:
Exercise:

Problem:

Suppose we change the original problem in [link] by using a 95% confidence level. Find a 95% confidence interval for the true (population) mean statistics exam score.

Solution:

To find the confidence interval, you need the sample mean, x, and the *EBM*.

- $x = 68$
- $EBM = \left(z_{\frac{\alpha}{2}}\right)\left(\frac{\sigma}{\sqrt{n}}\right)$
- $\sigma = 3$; $n = 36$; The confidence level is 95% (CL = 0.95).

CL = 0.95 so $\alpha = 1 - \text{CL} = 1 - 0.95 = 0.05$

$\frac{\alpha}{2} = 0.025 \qquad z_{\frac{\alpha}{2}} = z_{0.025}$

The area to the right of $z_{0.025}$ is 0.025 and the area to the left of $z_{0.025}$ is 1 – 0.025 = 0.975.

$z_{\frac{\alpha}{2}} = z_{0.025} = 1.96$

Note that this value was obtained using invNorm(0.975,0,1) on the TI-83, 83+, or 84+ calculators. (This can also be found using appropriate commands on other calculators, using a computer, or using a probability table for the standard normal distribution.)

$EBM = (1.96)\left(\frac{3}{\sqrt{36}}\right) = 0.98$

$x - EBM = 68 - 0.98 = 67.02$

$x + EBM = 68 + 0.98 = 68.98$

Notice that the *EBM* is larger for a 95% confidence level in the original problem.

Interpretation

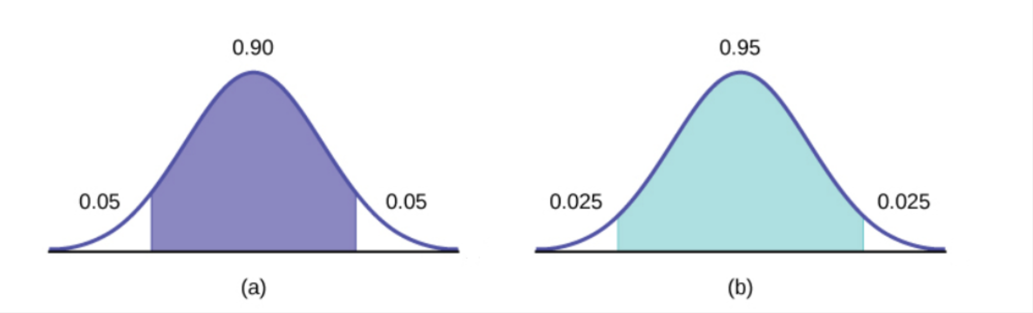
We estimate with 95% confidence that the true mean for all statistics exam scores is between 67.02 and 68.98.

Explanation of 95% Confidence Level

Ninety-five percent of all confidence intervals constructed in this way contain the true value of the population mean statistics exam score.

Comparing the results

The 90% confidence interval is (67.18, 68.82). The 95% confidence interval is (67.02, 68.98). The 95% confidence interval is wider. If you look at the graphs used to find the z-scores, because the area 0.95 is larger than the area 0.90, it makes sense that the 95% confidence interval is wider, since the z-score needed to give that larger central area must be larger, which in turn gives a larger calculated error bound. It should also be intuitive that in order to be more confident that the confidence interval actually does contain the true value of the population mean for all statistics exam scores, the confidence interval necessarily needs to be wider.



Summary: Effect of Changing the Confidence Level

- Increasing the confidence level increases the error bound, making the confidence interval wider.
- Decreasing the confidence level decreases the error bound, making the confidence interval narrower.

Note:

Try It

Exercise:

Problem:

Refer back to the pizza-delivery Try It exercise. The population standard deviation is six minutes and the sample mean deliver time is 36 minutes. Use a sample size of 20. Find a 95% confidence interval estimate for the true mean pizza delivery time.

Solution:

(33.37, 38.63)

Example:

Suppose we change the original problem in [link] to see what happens to the error bound if the sample size is changed.

Exercise:

Problem:

Leave everything the same except the sample size. Use the original 90% confidence level. What happens to the error bound and the confidence interval if we increase the sample size and use n = 100 instead of n = 36? What happens if we decrease the sample size to n = 25 instead of n = 36?

- $x = 68$
- $EBM = \left(z_{\frac{\alpha}{2}}\right)\left(\frac{\sigma}{\sqrt{n}}\right)$
- $\sigma = 3$; The confidence level is 90% (CL=0.90); $z_{\frac{\alpha}{2}} = z_{0.05} = 1.645$.

Solution:

Solution A

If we **increase** the sample size n to 100, we **decrease** the error bound.

When n = 100: $EBM = \left(z_{\frac{\alpha}{2}}\right)\left(\frac{\sigma}{\sqrt{n}}\right) = (1.645)\left(\frac{3}{\sqrt{100}}\right) = 0.4935$.

Solution:

Solution B
If we **decrease** the sample size n to 25, we **increase** the error bound.

When $n = 25$: $EBM = \left(z_{\frac{\alpha}{2}}\right)\left(\frac{\sigma}{\sqrt{n}}\right) = (1.645)\left(\frac{3}{\sqrt{25}}\right) = 0.987$.

Summary: Effect of Changing the Sample Size

- Increasing the sample size causes the error bound to decrease, making the confidence interval narrower.
- Decreasing the sample size causes the error bound to increase, making the confidence interval wider.

Note:
Try It
Exercise:

Problem:

Refer back to the pizza-delivery Try It exercise. The mean delivery time is 36 minutes and the population standard deviation is six minutes. Assume the sample size is changed to 50 restaurants with the same sample mean. Find a 90% confidence interval estimate for the population mean delivery time.

Solution:

(34.604, 37.396)

Working Backwards to Find the Error Bound or Sample Mean

When we calculate a confidence interval, we find the sample mean, calculate the error bound, and use them to calculate the confidence interval. However, sometimes when we read statistical studies, the study may state the confidence interval only. If we know the confidence interval, we can work backwards to find both the error bound and the sample mean.

Finding the Error Bound

- From the upper value for the interval, subtract the sample mean,
- OR, from the upper value for the interval, subtract the lower value. Then divide the difference by two.

Finding the Sample Mean

- Subtract the error bound from the upper value of the confidence interval,
- OR, average the upper and lower endpoints of the confidence interval.

Notice that there are two methods to perform each calculation. You can choose the method that is easier to use with the information you know.

Example:
Suppose we know that a confidence interval is **(67.18, 68.82)** and we want to find the error bound. We may know that the sample mean is 68, or perhaps our source only gave the confidence interval and did not tell us the value of the sample mean.

Calculate the Error Bound:

- If we know that the sample mean is 68: *EBM* = 68.82 – 68 = 0.82.
- If we don't know the sample mean: *EBM* = $\frac{(68.82-67.18)}{2}$ = 0.82.

Calculate the Sample Mean:

- If we know the error bound: x = 68.82 – 0.82 = 68
- If we don't know the error bound: x = $\frac{(67.18+68.82)}{2}$ = 68.

Note:
Try It

Exercise:

Problem:

Suppose we know that a confidence interval is (42.12, 47.88). Find the error bound and the sample mean.

Solution:

Sample mean is 45, error bound is 2.88

Calculating the Sample Size *n*

If researchers desire a specific margin of error, then they can use the error bound formula to calculate the required sample size.

The error bound formula for a population mean when the population standard deviation is known is
$EBM = \left(z_{\frac{\alpha}{2}}\right)\left(\frac{\sigma}{\sqrt{n}}\right)$.

The formula for sample size is $n = \frac{z^2\sigma^2}{EBM^2}$, found by solving the error bound formula for n.

In this formula, z is $z_{\frac{\alpha}{2}}$, corresponding to the desired confidence level. A researcher planning a study who wants a specified confidence level and error bound can use this formula to calculate the size of the sample needed for the study.

Example:
The population standard deviation for the age of Foothill College students is 15 years. If we want to be 95% confident that the sample mean age is within two years of the true population mean age of Foothill College students, how many randomly selected Foothill College students must be surveyed?

- From the problem, we know that $\sigma = 15$ and $EBM = 2$.
- $z = z_{0.025} = 1.96$, because the confidence level is 95%.

- $n = \frac{z^2\sigma^2}{EBM^2} = \frac{(1.96)^2(15)^2}{2^2}$ = 216.09 using the sample size equation.
- Use n = 217: Always round the answer UP to the next higher integer to ensure that the sample size is large enough.

Therefore, 217 Foothill College students should be surveyed in order to be 95% confident that we are within two years of the true population mean age of Foothill College students.

Note:
Try It
Exercise:

Problem:

The population standard deviation for the height of high school basketball players is three inches. If we want to be 95% confident that the sample mean height is within one inch of the true population mean height, how many randomly selected students must be surveyed?

Solution:

35 students

References

“American Fact Finder.” U.S. Census Bureau. Available online at http://factfinder2.census.gov/faces/nav/jsf/pages/searchresults.xhtml?refresh=t (accessed July 2, 2013).

“Disclosure Data Catalog: Candidate Summary Report 2012.” U.S. Federal Election Commission. Available online at http://www.fec.gov/data/index.jsp (accessed July 2, 2013).

“Headcount Enrollment Trends by Student Demographics Ten-Year Fall Trends to Most Recently Completed Fall.” Foothill De Anza Community College District. Available online at http://research.fhda.edu/factbook/FH_Demo_Trends/FoothillDemographicTrends.htm (accessed September 30,2013).

Kuczmarski, Robert J., Cynthia L. Ogden, Shumei S. Guo, Laurence M. Grummer-Strawn, Katherine M. Flegal, Zuguo Mei, Rong Wei, Lester R. Curtin, Alex F. Roche, Clifford L. Johnson. "2000 CDC Growth Charts for the United States: Methods and Development." Centers for Disease Control and Prevention. Available online at http://www.cdc.gov/growthcharts/2000growthchart-us.pdf (accessed July 2, 2013).

La, Lynn, Kent German. "Cell Phone Radiation Levels." c|net part of CBX Interactive Inc. Available online at http://reviews.cnet.com/cell-phone-radiation-levels/ (accessed July 2, 2013).

"Mean Income in the Past 12 Months (in 2011 Inflaction-Adjusted Dollars): 2011 American Community Survey 1-Year Estimates." American Fact Finder, U.S. Census Bureau. Available online at http://factfinder2.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_11_1YR_S1902&prodType=table (accessed July 2, 2013).

"Metadata Description of Candidate Summary File." U.S. Federal Election Commission. Available online at http://www.fec.gov/finance/disclosure/metadata/metadataforcandidatesummary.shtml (accessed July 2, 2013).

"National Health and Nutrition Examination Survey." Centers for Disease Control and Prevention. Available online at http://www.cdc.gov/nchs/nhanes.htm (accessed July 2, 2013).

Chapter Review

In this module, we learned how to calculate the confidence interval for a single population mean where the population standard deviation is known. When estimating a population mean, the margin of error is called the error bound for a population mean (*EBM*). A confidence interval has the general form:

(lower bound, upper bound) = (point estimate – *EBM*, point estimate + *EBM*) = ($x - EBM, x + EBM$)

The calculation of *EBM* depends on the size of the sample and the level of confidence desired. The confidence level is the percent of all possible samples that can be expected to include the true population parameter. As the confidence level increases, the corresponding *EBM* increases as well. As the sample size increases, the *EBM* decreases. By the central limit theorem,

$$EBM = \left(z_{\frac{\alpha}{2}}\right)\left(\frac{\sigma}{\sqrt{n}}\right)$$

Given a confidence interval, you can work backwards to find the error bound (*EBM*) or the sample mean. To find the error bound, find the difference of the upper bound of the interval and the mean. If you do not know the sample mean, you can find the error bound by calculating half the difference of the upper and lower bounds. To find the sample mean given a confidence interval, find the difference of the upper bound and the error bound. If the error bound is unknown, then average the upper and lower bounds of the confidence interval to find the sample mean.

Sometimes researchers know in advance that they want to estimate a population mean within a specific margin of error for a given level of confidence. In that case, solve the *EBM* formula for n to discover the size of the sample that is needed to achieve this goal:

$$n = \frac{z^2\sigma^2}{EBM^2}$$

In this formula, z is $z_{\frac{\alpha}{2}}$, corresponding to the desired confidence level.

Formula Review

$X \sim N\left(\mu_x, \frac{\sigma}{\sqrt{n}}\right)$ The distribution of sample means is normally distributed with mean equal to the population mean and standard deviation given by the population standard deviation divided by the square root of the sample size.

The general form for a confidence interval for a single population mean, known standard deviation, normal distribution is given by

(lower bound, upper bound) = (sample mean – *EBM*, sample mean + *EBM*)

= $(x - EBM, x + EBM)$

= $\left(x - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}, x + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right)$

EBM = $z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$ = the error bound for the mean, or the margin of error for a single population mean; this formula is used when the population standard deviation is known.

CL = confidence level, or the proportion of confidence intervals created that are expected to contain the true population parameter

α = 1 – CL = the proportion of confidence intervals that will not contain the population parameter

$z_{\frac{\alpha}{2}}$ = the z-score with the property that the area to the right of the z-score is $\frac{\alpha}{2}$ this is the z-score used in the calculation of *EBM* where α = 1 – CL.

n = $\frac{z^2\sigma^2}{EBM^2}$ = the formula used to determine the sample size (n) needed to achieve a desired margin of error at a given level of confidence. In this formula, z is $z_{\frac{\alpha}{2}}$, corresponding to the desired confidence level.

General form of a confidence interval

(lower value, upper value) = (point estimate − error bound, point estimate + error bound)

To find the error bound when you know the confidence interval

error bound = upper value−point estimate OR error bound = $\frac{\text{upper value}-\text{lower value}}{2}$

Single Population Mean, Known Standard Deviation, Normal Distribution

Use the Normal Distribution for Means, Population Standard Deviation is Known: *EBM* = $z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$

The confidence interval to estimate a population mean has the format (x − *EBM*, x + *EBM*).

Use the following information to answer the next five exercises: The standard deviation of the weights of elephants is known to be approximately 15 pounds. We wish to construct a 95% confidence interval for the mean weight of newborn elephant calves. Fifty newborn elephants are weighed. The sample mean is 244 pounds. The sample standard deviation is 11 pounds.

Exercise:

Problem: Identify the following:

a. x = ______
b. σ = ______
c. n = ______

Solution:

a. 244
b. 15
c. 50

Exercise:

Problem: In words, define the random variables X and X.

Exercise:

Problem: Which distribution should you use for this problem?

Solution:

$$N\left(244, \frac{15}{\sqrt{50}}\right)$$

Exercise:

Problem:

Construct a 95% confidence interval for the population mean weight of newborn elephants. State the confidence interval, and calculate the error bound.

Exercise:

Problem:

What will happen to the confidence interval obtained (with the same sample mean), if 500 newborn elephants are weighed instead of 50? Why?

Solution:

As the sample size increases, the error bound decreases, so the interval size decreases.

Use the following information to answer the next seven exercises: The U.S. Census Bureau conducts a study to determine the time needed to complete the short form. The Bureau surveys 200 people. The sample mean is 8.2 minutes. There is a known standard deviation of 2.2 minutes. The population distribution is assumed to be normal.

Exercise:

Problem: Identify the following:

a. x = ______
b. σ = ______
c. n = ______

Exercise:

Problem: In words, define the random variables X and X.

Solution:

X is the time in minutes it takes to complete the U.S. Census short form. X is the mean time it took a sample of 200 people to complete the U.S. Census short form.

Exercise:

Problem: Which distribution should you use for this problem?

Exercise:

Problem:

Construct a 90% confidence interval for the population mean time to complete the forms. State the confidence interval and calculate the error bound.

Solution:

CI: (7.9441, 8.4559)

EBM = 0.26

Exercise:

Problem:

If the Census wants to increase its level of confidence and keep the error bound the same by taking another survey, what changes should it make?

Exercise:

Problem:

If the Census did another survey, kept the error bound the same, and surveyed only 50 people instead of 200, what would happen to the level of confidence? Why?

Solution:

The level of confidence would decrease because decreasing n normally makes the confidence interval wider, so if the error bound stays the same, the confidence level must decrease.

Exercise:

Problem:

Suppose the Census needed to be 98% confident of the population mean length of time. Would the Census have to survey more people? Why or why not?

Use the following information to answer the next ten exercises: A sample of 20 heads of lettuce was selected. Assume that the population distribution of head weight is normal. The weight of each head of lettuce was then recorded. The mean weight was 2.2 pounds with a standard deviation of 0.1 pounds. The population standard deviation is known to be 0.2 pounds.

Exercise:

Problem: Identify the following:

a. x = _______
b. σ = _______
c. n = _______

Solution:

a. x = 2.2
b. σ = 0.2
c. n = 20

Exercise:

Problem: In words, define the random variable X.

Exercise:

Problem: In words, define the random variable X.

Solution:

X is the mean weight of a sample of 20 heads of lettuce.

Exercise:

Problem: Which distribution should you use for this problem?

Exercise:

Problem:

Construct a 90% confidence interval for the population mean weight of the heads of lettuce. State the confidence interval and calculate the error bound.

Solution:

$EBM = 0.07$
CI: (2.13, 2.27) rounded to two decimal places

Exercise:

Problem:

Construct a 95% confidence interval for the population mean weight of the heads of lettuce. State the confidence interval and calculate the error bound.

Exercise:

Problem:

In complete sentences, explain why the confidence interval in [link] is wider than in [link].

Solution:

The interval is greater because the level of confidence increased. If the only change made in the analysis is a change in confidence level, then all we are doing is changing how much area is being calculated for the normal distribution. Therefore, a larger confidence level results in larger areas and larger intervals.

Exercise:

Problem:

In complete sentences, give an interpretation of what the interval in [link] means.

Exercise:

Problem:

What would happen if 40 heads of lettuce were sampled instead of 20, and the error bound remained the same?

Solution:

The confidence level would increase.

Exercise:

Problem:

What would happen if 40 heads of lettuce were sampled instead of 20, and the confidence level remained the same?

Use the following information to answer the next 14 exercises: The mean age for all Foothill College students for a recent Fall term was 33.2. The population standard deviation has been pretty consistent at 15. Suppose that twenty-five Winter students were randomly selected. The mean age for the sample was 30.4. We are interested in the true mean age for Winter Foothill College students. Let X = the age of a Winter Foothill College student.

Exercise:

Problem: x = _____

Solution:

30.4

Exercise:

Problem: n = _____

Exercise:

Problem: ________ = 15

Solution:

σ

Exercise:

Problem: In words, define the random variable X.

Exercise:

Problem: What is $\bar{x}$ estimating?

Solution:

μ, the true mean age for Winter Foothill College students

Exercise:

Problem: Is σ_x known?

Exercise:

Problem:

As a result of your answer to [link], state the exact distribution to use when calculating the confidence interval.

Solution:

$N\left(30.4, \frac{15}{\sqrt{25}}\right)$

Construct a 95% Confidence Interval for the true mean age of Winter Foothill College students by working out then answering the next seven exercises.

Exercise:

Problem: How much area is in both tails (combined)? α =________

Exercise:

Problem: How much area is in each tail? $\frac{\alpha}{2}$ =________

Solution:

0.025

Exercise:

Problem: Identify the following specifications:

a. lower limit
b. upper limit
c. error bound

Exercise:

Problem: The 95% confidence interval is:____________________________.

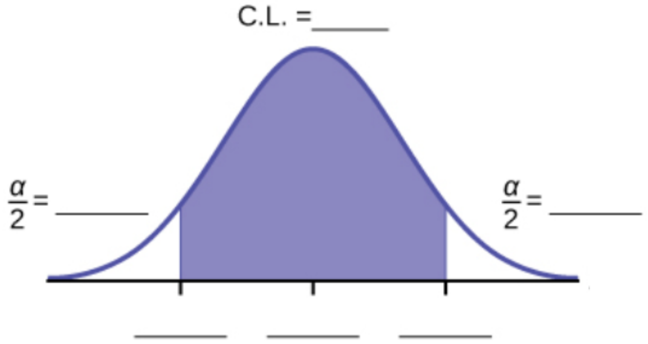
Solution:

(24.52,36.28)

Exercise:

Problem:

Fill in the blanks on the graph with the areas and upper and lower **z**-scores associated with the 95% confidence level, as well as the mean of this distribution.


C.L. =_____
$\frac{\alpha}{2}$ = _____
$\frac{\alpha}{2}$ = _____
_____ _____ _____

Exercise:

Problem: In one complete sentence, explain what the interval means.

Solution:

We are 95% confident that the true mean age for Winger Foothill College students is between 24.52 and 36.28 years.

Exercise:

Problem:

Using the same mean, standard deviation, and level of confidence, suppose that n were 69 instead of 25. Would the error bound become larger or smaller? How do you know?

Exercise:

Problem:

Using the same mean, standard deviation, and sample size, how would the error bound change if the confidence level were reduced to 90%? Why?

Solution:

The error bound for the mean would decrease because as the CL decreases, you need less area under the normal curve (which translates into a smaller interval) to capture the true population mean.

Homework

Exercise:

Problem:

Among various ethnic groups, the standard deviation of heights is known to be approximately three inches. We wish to construct a 95% confidence interval for the mean height of male Swedes. Forty-eight male Swedes are surveyed. The sample mean is 71 inches. The sample standard deviation is 2.8 inches.

a.
   i. x =_________
   ii. σ =_________
   iii. n =_________

b. In words, define the random variables X and X.
c. Which distribution should you use for this problem? Explain your choice.
d. Construct a 95% confidence interval for the population mean height of male Swedes.

    i. State the confidence interval.
    ii. Calculate the error bound.

e. What will happen to the level of confidence obtained if 1,000 male Swedes are surveyed instead of 48? Why?

Solution:

a.
    i. 71 inches
    ii. 3 inches
    iii. 48

b. X is the height of a Swedish male, and is the mean height from a sample of 48 male Swedes.
c. Normal. We know the standard deviation for the population, and the sample size is greater than 30.

d.
    i. CI: (70.151, 71.849)
    ii. *EBM* = 0.849

e. The confidence interval will decrease in size, because the sample size increased. Recall, when all other factors remain unchanged, an increase in sample size decreases variability. Thus, we do not need as large an interval to capture the true population mean.

Exercise:

Problem:

Announcements for 84 upcoming engineering conferences were randomly picked from a stack of IEEE Spectrum magazines. The mean length of the conferences was 3.94 days. Assume the underlying population is normal, with a standard deviation of 1.28 days.

a. In words, define the random variables X and X.
b. Which distribution should you use for this problem? Explain your choice.
c. Construct a 95% confidence interval for the population mean length of engineering conferences.

    i. State the confidence interval.
    ii. Calculate the error bound.

Exercise:

Problem:

Suppose that an accounting firm does a study to determine the time needed to complete one person’s tax forms. It randomly surveys 100 people. The sample mean is 23.6 hours. There is a known population standard deviation of 7.0 hours. The population distribution is assumed to be normal.

a.
    i. x =________
    ii. σ =________
    iii. n =________

b. In words, define the random variables X and X.
c. Which distribution should you use for this problem? Explain your choice.
d. Construct a 90% confidence interval for the population mean time to complete the tax forms.
    i. State the confidence interval.
    ii. Calculate the error bound.
e. If the firm wished to increase its level of confidence and keep the error bound the same by taking another survey, what changes should it make?
f. If the firm did another survey, kept the error bound the same, and only surveyed 49 people, what would happen to the level of confidence? Why?
g. Suppose that the firm decided that it needed to be at least 96% confident of the population mean length of time to within one hour. How would the number of people the firm surveys change? Why?

Solution:

a.
    i. x = 23.6 hours
    ii. σ = 7 hours
    iii. n = 100

b. X is the time needed to complete an individual tax form. X is the mean time to complete tax forms from a sample of 100 customers.
c. $N\left(23.6, \frac{7}{\sqrt{100}}\right)$ because we are estimating the mean and we know the population standard deviation.

d. i. (22.449, 24.751)
   ii. *EBM* = 1.151

e. It will need to change the sample size. The firm needs to determine what the confidence level should be, then apply the error bound formula to determine the necessary sample size.
f. The confidence level would increase as a result of a larger interval. Smaller sample sizes result in more variability. To capture the true population mean, we need to have a larger interval.
g. According to the error bound formula, the firm needs to survey 206 people. Since we increase the confidence level, we need to increase either our error bound or the sample size.

Exercise:

Problem:

A sample of 16 small bags of the same brand of candies was selected. Assume that the population distribution of bag weights is normal. The weight of each bag was then recorded. The mean weight was two ounces with a standard deviation of 0.12 ounces. The population standard deviation is known to be 0.1 ounce.

a. i. x =________
   ii. σ =________
   iii. s_{x} =________

b. In words, define the random variable X.
c. In words, define the random variable X.
d. Which distribution should you use for this problem? Explain your choice.
e. Construct a 90% confidence interval for the population mean weight of the candies.

   i. State the confidence interval.

    ii. Calculate the error bound.

f. Construct a 98% confidence interval for the population mean weight of the candies.

    i. State the confidence interval.
    ii. Calculate the error bound.

g. In complete sentences, explain why the confidence interval in part f is larger than the confidence interval in part e.
h. In complete sentences, give an interpretation of what the interval in part f means.

Exercise:

Problem:

A camp director is interested in the mean number of letters each child sends during his or her camp session. The population standard deviation is known to be 2.5. A survey of 20 campers is taken. The mean from the sample is 7.9 with a sample standard deviation of 2.8. Assume the number of letters each child sends is normally distributed.

a.
    i. x =________
    ii. σ =________
    iii. n =________

b. Define the random variables X and X in words.
c. Which distribution should you use for this problem? Explain your choice.
d. Construct a 90% confidence interval for the population mean number of letters campers send home.

    i. State the confidence interval.
    ii. Calculate the error bound.

e. What will happen to the error bound and confidence interval if 500 campers are surveyed? Why?

Solution:

a.
   i. 7.9
   ii. 2.5
   iii. 20

b. X is the number of letters a single camper will send home. X is the mean number of letters sent home from a sample of 20 campers.

c. $N\left(7.9, \frac{2.5}{\sqrt{20}}\right)$

d.
   i. CI: (6.98, 8.82)
   ii. *EBM*: 0.92

e. The error bound and confidence interval will decrease.

Exercise:

Problem:

What is meant by the term "90% confident" when constructing a confidence interval for a mean?

a. If we took repeated samples, approximately 90% of the samples would produce the same confidence interval.
b. If we took repeated samples, approximately 90% of the confidence intervals calculated from those samples would contain the sample mean.
c. If we took repeated samples, approximately 90% of the confidence intervals calculated from those samples would contain the true value of the population mean.
d. If we took repeated samples, the sample mean would equal the population mean in approximately 90% of the samples.

Exercise:

Problem:

The Federal Election Commission collects information about campaign contributions and disbursements for candidates and political committees each election cycle. During the 2012 campaign season, there were 1,619 candidates for the House of Representatives across the United States who received contributions from individuals. The following table shows the total receipts from individuals for a random selection of 40 House candidates rounded to the nearest $100. The standard deviation for this data to the nearest hundred is σ = $909,200.

$3,600	$1,243,900	$10,900	$385,200	$581,500
$7,400	$2,900	$400	$3,714,500	$632,500
$391,000	$467,400	$56,800	$5,800	$405,200
$733,200	$8,000	$468,700	$75,200	$41,000
$13,300	$9,500	$953,800	$1,113,500	$1,109,300
$353,900	$986,100	$88,600	$378,200	$13,200
$3,800	$745,100	$5,800	$3,072,100	$1,626,700
$512,900	$2,309,200	$6,600	$202,400	$15,800

a. Find the point estimate for the population mean.
b. Using 95% confidence, calculate the error bound.
c. Create a 95% confidence interval for the mean total individual contributions.
d. Interpret the confidence interval in the context of the problem.

Solution:

a. x = \$568,873
b. CL = 0.95 α = 1 – 0.95 = 0.05 $z_{\frac{\alpha}{2}}$ = 1.96

$$EBM = z_{0.025}\frac{\sigma}{\sqrt{n}} = 1.96\ \frac{909200}{\sqrt{40}} = \$281{,}764$$

c. $x - EBM = 568{,}873 - 281{,}764 = 287{,}109$
$x + EBM = 568{,}873 + 281{,}764 = 850{,}637$

Alternate solution:

Note:

1. Press `STAT` and arrow over to `TESTS`.
2. Arrow down to `7:ZInterval`.
3. Press `ENTER`.
4. Arrow to Stats and press `ENTER`.
5. Arrow down and enter the following values:
    - σ : 909,200
    - x: 568,873
    - n: 40
    - CL: 0.95
6. Arrow down to Calculate and press `ENTER`.
7. The confidence interval is (\$287,114, \$850,632).
8. Notice the small difference between the two solutions–these differences are simply due to rounding error in the hand calculations.

d. We estimate with 95% confidence that the mean amount of contributions received from all individuals by House candidates is between \$287,109 and \$850,637.

Exercise:

Problem:

The American Community Survey (ACS), part of the United States Census Bureau, conducts a yearly census similar to the one taken every ten years, but with a smaller percentage of participants. The most recent survey estimates with 90% confidence that the mean household income in the U.S. falls between \$69,720 and \$69,922. Find the point estimate for mean U.S. household income and the error bound for mean U.S. household income.

Exercise:

Problem:

The average height of young adult males has a normal distribution with standard deviation of 2.5 inches. You want to estimate the mean height of students at your college or university to within one inch with 93% confidence. How many male students must you measure?

Solution:

Use the formula for *EBM*, solved for n:
$n = \frac{z^2\sigma^2}{EBM^2}$

From the statement of the problem, you know that σ = 2.5, and you need *EBM* = 1.

$z = z_{0.035} = 1.812$

(This is the value of z for which the area under the normal curve to the ***right*** of z is 0.035.)

$$n = \frac{z^2\sigma^2}{EBM^2} = \frac{1.812^2 2.5^2}{1^2} \approx 20.52$$

You need to measure at least 21 male students to achieve your goal.

Glossary

Confidence Level (CL)

the percent expression for how confident one can be that the confidence interval contains the true population parameter because it represents the percent of all possible confidence intervals constructed, using the method taught in this section, that will contain the true population parameter, in repeated sampling; for example, if the CL = 90%, then in 90 out of 100 samples the confidence interval that is constructed will enclose the true population parameter.

Error Bound for a Population Mean (*EBM*)

the margin of error when estimating a single population mean; depends on the confidence level, sample size, and known or estimated population standard deviation.

Estimating a Single Population Mean using the Student t-Distribution

In practice, we rarely know the population **standard deviation**. In the past, when the sample size was large, this did not present a problem to statisticians. They used the sample standard deviation s as an estimate for σ and proceeded as before to calculate a **confidence interval** with close enough results. However, statisticians ran into problems when the sample size was small. A small sample size caused inaccuracies in the confidence interval.

William S. Goset (1876–1937) of the Guinness brewery in Dublin, Ireland ran into this problem. His experiments with hops and barley produced very few samples. Just replacing σ with s did not produce accurate results when he tried to calculate a confidence interval. He realized that he could not use a normal distribution for the calculation; he found that the actual distribution depends on the sample size. This problem led him to "discover" what is called the **Student's t-distribution**. The name comes from the fact that Gosset wrote under the pen name "Student."

Up until the mid-1970s, some statisticians used the **normal distribution** approximation for large sample sizes and only used the Student's t-distribution for sample sizes of at most 30. With graphing calculators and computers, the practice now is to use the Student's t-distribution whenever s is used as an estimate for σ.

If you draw a simple random sample of size n from a population that has an approximately normal distribution with mean μ and unknown population standard deviation σ and calculate the t-score: $t = \frac{x-\mu}{\left(\frac{s}{\sqrt{n}}\right)}$, then the t-scores follow a **Student's t-distribution with n – 1 degrees of freedom**. The t-score has the same interpretation as the **z-score**. It measures how far x is from its mean μ. For each sample size n, there is a different Student's t-distribution.

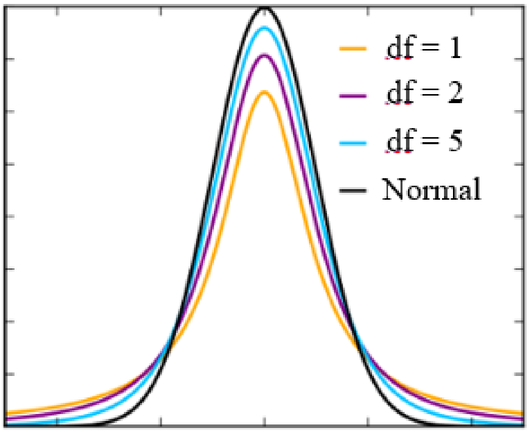
The **degrees of freedom**, $n-1$, come from the calculation of the sample standard deviation s. In Measuring the Spread of Data, we used n deviations $(x-x\text{values})$ to calculate s. Because the sum of the deviations is zero, we can find the last deviation once we know the other $n-1$ deviations. The other $n-1$ deviations can change or vary freely. **We call the number $n-1$ the degrees of freedom (df).**

Properties of the Student's t-Distribution

- The graph for the Student's t-distribution is similar to the standard normal curve.
- The mean for the Student's t-distribution is zero and the distribution is symmetric about zero.
- The Student's t-distribution has more probability in its tails than the standard normal distribution because the spread of the t-distribution is greater than the spread of the standard normal. So the graph of the Student's t-distribution will be thicker in the tails and shorter in the center than the graph of the standard normal distribution.

- The exact shape of the Student's t-distribution depends on the degrees of freedom. As the degrees of freedom increases, the graph of Student's t-distribution becomes more like the graph of the standard normal distribution.
- The underlying population of individual observations is assumed to be normally distributed with unknown population mean μ and unknown population standard deviation σ. The size of the underlying population is generally not relevant unless it is very small. If it is bell shaped (normal) then the assumption is met and doesn't need discussion. Random sampling is assumed, but that is a completely separate assumption from normality.

The figure below compares various Student's t-distributions and the standard normal distribution.



Calculators and computers can easily calculate any Student's t-probabilities. The TI-83,83+, and 84+ have a tcdf function to find the probability for given values of t. The format for the tcdf command is tcdf(lower bound, upper bound, degrees of freedom). However, for confidence intervals, we need to use **inverse** probability to find the value of t when we know the probability.

For the TI-84+ you can use the invT command on the DISTRibution menu. The invT command works similarly to the invnorm. The invT command requires two inputs: **invT(area to the left, degrees of freedom)** The output is the t-score that corresponds to the area we specified.

Note: The TI-83 and 83+ do not have the invT command.

A probability table for the Student's t-distribution can also be used. The table gives t-scores that correspond to the confidence level (column) and degrees of freedom (row). (The TI-83/83+ does not have an invT command, so if you are using that calculator, you need to use a probability table for the Student's t-Distribution when calculating a confidence interval by hand.) When using a t-table, note that some tables are formatted to show the confidence level in the column headings, while the column headings in some tables may show only corresponding area in one or both tails.

A Student's t table (See Appendix B) gives t-scores given the degrees of freedom and the right-tailed probability. The table is very limited. **Calculators and computers can easily calculate any Student's t-probabilities.**

The notation for the Student's t-distribution (using *T* as the random variable) is:

- $T \sim t_{\text{df}}$ where df = $n - 1$.
- For example, if we have a sample of size n = 20 items, then we calculate the degrees of freedom as df = $n - 1$ = 20 - 1 = 19 and we write the distribution as $T \sim t_{19}$.

If the population standard deviation is not known, the **error bound for a population mean** is:

- $EBM = t_{\frac{\alpha}{2}}\left(\frac{s}{\sqrt{n}}\right)$
- $t_{\frac{\alpha}{2}}$ is the t-score with area to the right equal to $\frac{\alpha}{2}$,
- use df = $n - 1$ degrees of freedom, and
- s = sample standard deviation.

The format for the confidence interval is:
$(x - EBM, x + EBM)$.

Note:
To calculate the confidence interval directly on a TI 83/83+/84+:
Press STAT.
Arrow over to TESTS.
Arrow down to 8:TInterval and press ENTER (or just press 8).

Example:
Exercise:

Problem:

Suppose you do a study of acupuncture to determine how effective it is in relieving pain. You measure sensory rates for 15 subjects with the results given. Use the sample data to construct a 95% confidence interval for the mean sensory rate for the population (assumed normal) from which you took the data.
The solution is shown step-by-step and by using the TI-83, 83+, or 84+ calculators.

8.6 9.4 7.9 6.8 8.3 7.3 9.2 9.6 8.7 11.4 10.3 5.4 8.1 5.5 6.9

- The first solution is step-by-step (Solution A).
- The second solution uses the TI-83+ and TI-84 calculators (Solution B).

Solution:

To find the confidence interval, you need the sample mean, x, and the *EBM*.

x = 8.2267; s = 1.6722; n = 15

df = 15 – 1 = 14; CL = 0.95, so α = 1 – CL = 1 – 0.95 = 0.05

$\frac{\alpha}{2}$ = 0.025; So $t_{\frac{\alpha}{2}} = t_{0.025}$

The area to the right of $t_{0.025}$ is 0.025, and the area to the left of $t_{0.025}$ is 1 – 0.025 = 0.975

$t_{\frac{\alpha}{2}} = t_{0.025} = 2.14$, using invT(.975,14) on the TI-84+ calculator.

$$EBM = t_{\frac{\alpha}{2}}\left(\frac{s}{\sqrt{n}}\right)$$

$$EBM = (2.14)\left(\frac{1.6722}{\sqrt{15}}\right) = 0.924$$

x – *EBM* = 8.2267 – 0.9240 = 7.30

x + *EBM* = 8.2267 + 0.9240 = 9.15

The 95% confidence interval is (7.30, 9.15).

We estimate with 95% confidence that the true population mean sensory rate is between 7.30 and 9.15.

Note:
Note
When calculating the error bound, a probability table for the Student's t-distribution can also be used to find the value of t. The table gives t-scores that correspond to the confidence level (column) and degrees of freedom (row); the t-score is found where the row and column intersect in the table.

Solution:

Note:
(After you've entered the data into a list.) Press `STAT` and arrow over to `TESTS`.
Arrow down to `8:TInterval` and press `ENTER` (or you can just press `8`).
Arrow to `Data` and press `ENTER`.
Arrow down to `List` and enter the list name where you put the data.
There should be a 1 after `Freq`.
Arrow down to `C-level` and enter 0.95
Arrow down to `Calculate` and press `ENTER`.
The 95% confidence interval is (7.3006, 9.1527)

Note:
Try It
Exercise:

Problem:

You do a study of hypnotherapy to determine how effective it is in increasing the number of hours of sleep subjects get each night. You measure hours of sleep for 12 subjects with the following results. Construct a 95% confidence interval for the mean number of hours slept for the population (assumed normal) from which you took the data.

8.2; 9.1; 7.7; 8.6; 6.9; 11.2; 10.1; 9.9; 8.9; 9.2; 7.5; 10.5

Solution:

(8.1634, 9.8032)

Example:
Exercise:

Problem:

The Human Toxome Project (HTP) is working to understand the scope of industrial pollution in the human body. Industrial chemicals may enter the body through pollution or as ingredients in consumer products. In October 2008, the scientists at HTP tested cord blood samples for 20 newborn infants in the United States. The cord blood of the "In utero/newborn" group was tested for 430 industrial compounds, pollutants, and other chemicals, including chemicals linked to brain and nervous system toxicity, immune system toxicity, and reproductive toxicity, and fertility problems. There are health concerns about the effects of some chemicals on the brain and nervous system. The following table shows how many of the targeted chemicals were found in each infant's cord blood.

79	145	147	160	116	100	159	151	156	126
137	83	156	94	121	144	123	114	139	99

Use this sample data to construct a 90% confidence interval for the mean number of targeted industrial chemicals to be found in an in infant's blood.

Solution:

Solution A
From the sample, you can calculate x = 127.45 and s = 25.965. There are 20 infants in the sample, so n = 20, and df = 20 – 1 = 19.

You are asked to calculate a 90% confidence interval: CL = 0.90, so α = 1 – CL = 1 – 0.90 = 0.10 $\frac{\alpha}{2} = 0.05, t_{\frac{\alpha}{2}} = t_{0.05}$

By definition, the area to the right of $t_{0.05}$ is 0.05 and so the area to the left of $t_{0.05}$ is 1 – 0.05 = 0.95.

Use a table, calculator, or computer to find that $t_{0.05}$ = 1.729.

$$EBM = t_{\frac{\alpha}{2}}\left(\frac{s}{\sqrt{n}}\right) = 1.729\left(\frac{25.965}{\sqrt{20}}\right) \approx 10.038$$

x – *EBM* = 127.45 – 10.038 = 117.412

x + *EBM* = 127.45 + 10.038 = 137.488

We estimate with 90% confidence that the mean number of all targeted industrial chemicals found in cord blood in the United States is between 117.412 and 137.488.

Solution:

Solution B

Note:
Enter the data as a list.
Press `STAT` and arrow over to `TESTS`.
Arrow down to `8:TInterval` and press `ENTER` (or you can just press `8`). Arrow to Data and press `ENTER`.
Arrow down to `List` and enter the list name where you put the data.
Arrow down to `Freq` and enter 1.
Arrow down to `C-level` and enter 0.90
Arrow down to `Calculate` and press `ENTER`.
The 90% confidence interval is (117.41, 137.49).

Note:
Try It
Exercise:

Problem:

A random sample of statistics students were asked to estimate the total number of hours they spend watching television in an average week. The responses are recorded in the following table. Use this sample data to construct a 98% confidence interval for the mean number of hours statistics students will spend watching television in one week.

0	3	1	20	9
5	10	1	10	4
14	2	4	4	5

Solution:
Solution A

x= 6.133, s = 5.514, n = 15, and df = 15 – 1 = 14

CL = 0.98, so α = 1 - CL = 1 - 0.98 = 0.02

$\frac{\alpha}{2} = 0.01; t_{\frac{\alpha}{2}} = t_{0.01} = 2.624$

$EBM = t_{\frac{\alpha}{2}} \left(\frac{s}{\sqrt{n}}\right) = 2.624 \left(\frac{5.514}{\sqrt{15}}\right) = 3.736$

x – EBM = 6.133 – 3.736 = 2.397

x + EBM = 6.133 + 3.736 = 9.869

We estimate with 98% confidence that the mean number of hours that statistics students spend watching television in one week is between 2.397 and 9.869.

Solution:
Solution B

Note:
Enter the data as a list.
Press `STAT` and arrow over to `TESTS`.
Arrow down to `8:TInterval`.
Press `ENTER`.
Arrow to `Data` and press `ENTER`.
Arrow down and enter the name of the list where the data is stored.
Enter `Freq`: 1
Enter `C-Level`: 0.98
Arrow down to `Calculate` and press `Enter`.
The 98% confidence interval is (2.3965, 9.8702).

References

“America’s Best Small Companies.” Forbes, 2013. Available online at http://www.forbes.com/best-small-companies/list/ (accessed July 2, 2013).

Data from *Microsoft Bookshelf*.

Data from http://www.businessweek.com/.

Data from http://www.forbes.com/.

"Disclosure Data Catalog: Leadership PAC and Sponsors Report, 2012." Federal Election Commission. Available online at http://www.fec.gov/data/index.jsp (accessed July 2,2013).

"Human Toxome Project: Mapping the Pollution in People." Environmental Working Group. Available online at http://www.ewg.org/sites/humantoxome/participants/participant-group.php?group=in+utero%2Fnewborn (accessed July 2, 2013).

"Metadata Description of Leadership PAC List." Federal Election Commission. Available online at http://www.fec.gov/finance/disclosure/metadata/metadataLeadershipPacList.shtml (accessed July 2, 2013).

Section Review

In many cases, the researcher does not know the population standard deviation, σ, of the measure being studied. In these cases, it is common to use the sample standard deviation, s, as an estimate of σ. The normal distribution creates accurate confidence intervals when σ is known, but it is not as accurate when s is used as an estimate. In this case, the Student's t-distribution is much better. Define a t-score using the following formula:

$$t = \frac{x - \ \mu}{{}^{s}/_{\sqrt{n}}}$$

The t-score follows the Student's t-distribution with $n{-}1$ degrees of freedom. The confidence interval under this distribution is calculated with $EBM = t_{\frac{\alpha}{2}}\left(\frac{s}{\sqrt{n}}\right)$ where $t_{\frac{\alpha}{2}}$ is the t-score with area to the right equal to $\frac{\alpha}{2}$, s is the sample standard deviation, and n is the sample size. Use a table, calculator, or computer to find $t_{\frac{\alpha}{2}}$ for a given α.

Formula Review

s = the standard deviation of the sample data values.

$t = \frac{x-\mu}{\frac{s}{\sqrt{n}}}$ is the formula for the t-score which measures how far away a measure is from the population mean in the Student's t-distribution

df = $n - 1$; the degrees of freedom for a Student's t-distribution where n represents the size of the sample

$T \sim t_{\mathrm{df}}$; the random variable, T, has a Student's t-distribution with df degrees of freedom

$EBM = t_{\frac{\alpha}{2}}\left(\frac{s}{\sqrt{n}}\right)$ = the error bound for the population mean when the population standard deviation is unknown

$t_{\frac{\alpha}{2}}$ is the t-score in the Student’s t-distribution with area to the right equal to $\frac{\alpha}{2}$

The general form for a confidence interval for a single mean, population standard deviation unknown, Student's t is given by (lower bound, upper bound)

= (point estimate – *EBM*, point estimate + *EBM*)

= $\left(x - t_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}}, x + t_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}}\right)$

Use the following information to answer the next five exercises. A hospital is trying to cut down on emergency room wait times. It is interested in the amount of time patients must wait before being called back to be examined. An investigation committee randomly surveyed 70 patients. The sample mean was 1.5 hours with a sample standard deviation of 0.5 hours.

Exercise:

Problem: Identify the following:

a. x =________
b. s_x =________
c. n =________
d. n – 1 =________

Exercise:

Problem: Define the random variables X and X in words.

Solution:

X is the number of hours a patient waits in the emergency room before being called back to be examined. X is the mean wait time of 70 patients in the emergency room.

Exercise:

Problem: Which distribution should you use for this problem?

Exercise:

Problem:

Construct a 95% confidence interval for the population mean time spent waiting. State the confidence interval and calculate the error bound.

Solution:

CI: (1.3808, 1.6192)

EBM = 0.12

Exercise:

Problem: Explain in complete sentences what the confidence interval means.

Use the following information to answer the next six exercises: One hundred eight Americans were surveyed to determine the number of hours they spend watching television each month. It was revealed that they watched an average of 151 hours each month with a standard deviation of 32 hours. Assume that the underlying population distribution is normal.

Exercise:

Problem: Identify the following:

a. x =________
b. s_x =________
c. n =________
d. n – 1 =________

Solution:

a. x = 151
b. s_x = 32
c. n = 108
d. n – 1 = 107

Exercise:

Problem: Define the random variable X in words.

Exercise:

Problem: Define the random variable X in words.

Solution:

X is the mean number of hours spent watching television per month from a sample of 108 Americans.

Exercise:

Problem: Which distribution should you use for this problem?

Exercise:

Problem:

Construct a 99% confidence interval for the population mean hours spent watching television per month. (a) State the confidence interval and (b) calculate the error bound.

Solution:

CI: (142.92, 159.08)

EBM = 8.08

Exercise:

Problem:

Why would the error bound change if the confidence level were lowered to 95%?

Use the following information to answer the next 13 exercises: The data in the following table are the result of a random survey of 39 national flags (with replacement between picks) from various countries. We are interested in finding a confidence interval for the true mean number of colors on a national flag. Let X = the number of colors on a national flag.

X	**Freq.**
1	1
2	7
3	18
4	7
5	6

Exercise:

Problem: Calculate the following:

a. x =_______
b. s_x =_______

c. n =_______

Solution:

a. 3.26
b. 1.02
c. 39

Exercise:

Problem: Define the random variable X in words.

Exercise:

Problem: What is x estimating?

Solution:

μ

Exercise:

Problem: Is σ_x known?

Exercise:

Problem:

As a result of your answer to [link], state the exact distribution to use when calculating the confidence interval.

Solution:

t_{38}

Construct a 95% confidence interval for the true mean number of colors on national flags.

Exercise:

Problem: How much area is in both tails (combined)?

Exercise:

Problem: How much area is in each tail?

Solution:

0.025

Exercise:

Problem: Calculate the following:

a. lower limit
b. upper limit

Exercise:

Problem: The 95% confidence interval is_____.

Solution:

(2.93, 3.59)

Exercise:

Problem: The error bound is ______.

Exercise:

Problem: In one complete sentence, explain what the interval means.

Solution:

We are 95% confident that the true mean number of colors for national flags is between 2.93 colors and 3.59 colors.

Exercise:

Problem:

Using the same x, s_x, and level of confidence, suppose that n were 69 instead of 39. Would the error bound become larger or smaller? How do you know?

Solution:

The error bound would become EBM = 0.245. This error bound decreased because as sample sizes increase, variability decreases and we need less interval length to capture the true mean.

Exercise:

Problem:

Using the same x, s_x, and n = 39, how would the error bound change if the confidence level were reduced to 90%? Why?

Homework

Exercise:

Problem:

A random survey of enrollment at 35 community colleges across the United States yielded the following figures: 6,414; 1,550; 2,109; 9,350; 21,828; 4,300; 5,944; 5,722; 2,825; 2,044; 5,481; 5,200; 5,853; 2,750; 10,012; 6,357; 27,000; 9,414; 7,681; 3,200; 17,500; 9,200; 7,380; 18,314; 6,557; 13,713; 17,768; 7,493; 2,771; 2,861; 1,263; 7,285; 28,165; 5,080; 11,622. Assume the underlying population is normal.

a.
   i. x = ___________
   ii. s_x = ___________
   iii. n = ___________
   iv. n – 1 = ___________

b. Define the random variables X and X in words.
c. Which distribution should you use for this problem? Explain your choice.
d. Construct a 95% confidence interval for the population mean enrollment at community colleges in the United States.

   i. State the confidence interval.
   ii. Calculate the error bound.

e. What will happen to the error bound and confidence interval if 500 community colleges were surveyed? Why?

Solution:

a.
   i. 8629
   ii. 6944
   iii. 35
   iv. 34

b. t_{34}

c.
   i. CI: (6244, 11,014)
   ii. EB = 2385

d. It will become smaller.

Exercise:

Problem:

Suppose that a committee is studying whether or not there is waste of time in our judicial system. It is interested in the mean amount of time individuals waste at the courthouse waiting to be called for jury duty. The committee randomly surveyed 81 people who recently served as jurors. The sample mean wait time was eight hours with a sample standard deviation of four hours.

a. i. x = ___________
   ii. s_x = ___________
   iii. n = ___________
   iv. $n - 1$ = ___________

b. Define the random variables X and X in words.
c. Which distribution should you use for this problem? Explain your choice.
d. Construct a 95% confidence interval for the population mean time wasted.

   i. State the confidence interval.
   ii. Calculate the error bound.

e. Explain in a complete sentence what the confidence interval means.

Exercise:

Problem:

A pharmaceutical company makes tranquilizers. It is assumed that the distribution for the length of time they last is approximately normal. Researchers in a hospital used the drug on a random sample of nine patients. The effective period of the tranquilizer for each patient (in hours) was as follows: 2.7; 2.8; 3.0; 2.3; 2.3; 2.2; 2.8; 2.1; and 2.4.

a. i. x = ___________
   ii. s_x = ___________
   iii. n = ___________
   iv. $n - 1$ = ___________

b. Define the random variable X in words.
c. Define the random variable X in words.
d. Which distribution should you use for this problem? Explain your choice.
e. Construct a 95% confidence interval for the population mean length of time.

   i. State the confidence interval.
   ii. Calculate the error bound.

f. What does it mean to be "95% confident" in this problem?

Solution:

- a.
  - i. x = 2.51
  - ii. s_x = 0.318
  - iii. n = 9
  - iv. n – 1 = 8
- b. the effective length of time for a tranquilizer
- c. the mean effective length of time of tranquilizers from a sample of nine patients
- d. We need to use a Student’s-t distribution, because we do not know the population standard deviation.
- e.
  - i. CI: (2.27, 2.75)
  - ii. *EBM*: 0.24
- f. If we were to sample many groups of nine patients and constructed a confidence interval using this procedure for each sample, 95% of the confidence intervals would contain the true population mean length of time.

Exercise:

Problem:

Suppose that 14 children, who were learning to ride two-wheel bikes, were surveyed to determine how long they had to use training wheels. It was revealed that they used them an average of six months with a sample standard deviation of three months. Assume that the underlying population distribution is normal.

- a.
  - i. x = ___________
  - ii. s_x = ___________
  - iii. n = ___________
  - iv. n – 1 = ___________
- b. Define the random variable X in words.
- c. Define the random variable X in words.
- d. Which distribution should you use for this problem? Explain your choice.
- e. Construct a 99% confidence interval for the population mean length of time using training wheels.
  - i. State the confidence interval.
  - ii. Calculate the error bound.
- f. Why would the error bound change if the confidence level were lowered to 90%?

Exercise:

Problem:

The Federal Election Commission (FEC) collects information about campaign contributions and disbursements for candidates and political committees each election cycle. A political action committee (PAC) is a committee formed to raise money for candidates and campaigns. A Leadership PAC is a PAC formed by a federal politician (senator or representative) to raise money to help other candidates' campaigns.

The FEC has reported financial information for 556 Leadership PACs that operating during the 2011–2012 election cycle. The following table shows the total receipts during this cycle for a random selection of 30 Leadership PACs.

\$46,500.00	\$0	\$40,966.50	\$105,887.20	\$5,175.00
\$29,050.00	\$19,500.00	\$181,557.20	\$31,500.00	\$149,970.80
\$2,555,363.20	\$12,025.00	\$409,000.00	\$60,521.70	\$18,000.00
\$61,810.20	\$76,530.80	\$119,459.20	\$0	\$63,520.00
\$6,500.00	\$502,578.00	\$705,061.10	\$708,258.90	\$135,810.00
\$2,000.00	\$2,000.00	\$0	\$1,287,933.80	\$219,148.30

$x = \$251,854.23$

$s = \ \$521,130.41$

Use this sample data to construct a 96% confidence interval for the mean amount of money raised by all Leadership PACs during the 2011–2012 election cycle. Use the Student's t-distribution.

Solution:

$x = \$251,854.23$

$s = \ \$521,130.41$

Note that we are not given the population standard deviation, only the standard deviation of the sample.

There are 30 measures in the sample, so n = 30, and df = 30 - 1 = 29

CL = 0.96, so α = 1 - CL = 1 - 0.96 = 0.04

$\frac{\alpha}{2} = 0.02; t_{\frac{\alpha}{2}} = t_{0.02}$ = 2.150

$$EBM = t_{\frac{\alpha}{2}}\left(\frac{s}{\sqrt{n}}\right) = 2.150\left(\frac{521{,}130.41}{\sqrt{30}}\right) \ = \ \$204,561.66$$

x - EBM = \$251,854.23 - \$204,561.66 = \$47,292.57

x + EBM = \$251,854.23+ \$204,561.66 = \$456,415.89

We estimate with 96% confidence that the mean amount of money raised by all Leadership PACs during the 2011–2012 election cycle lies between \$47,292.57 and \$456,415.89.

Alternate Solution

Note:
Press `STAT` and arrow over to `TESTS`.
Arrow down to `8:TInterval`.
Press `ENTER`.
Arrow to Stats and press `ENTER`.
Enter x: 251854.23
Enter s_x: 521130.41
Enter `n`: 30
Enter `C-Level`: 0.96
Arrow down to `Calculate` and press `Enter`.
The 96% confidence interval is (\$47,262, \$456,447).

The difference between solutions arises from rounding differences.

Exercise:

Problem:

Forbes magazine published data on the best small firms in 2012. These were firms that had been publicly traded for at least a year, have a stock price of at least \$5 per share, and have reported annual revenue between \$5 million and \$1 billion. The following table shows the ages of the corporate CEOs for a random sample of these firms.

48	58	51	61	56
59	74	63	53	50
59	60	60	57	46
55	63	57	47	55
57	43	61	62	49
67	67	55	55	49

Use this sample data to construct a 90% confidence interval for the mean age of CEO’s for these top small firms. Use the Student's t-distribution.

Exercise:

Problem:

Unoccupied seats on flights cause airlines to lose revenue. Suppose a large airline wants to estimate its mean number of unoccupied seats per flight over the past year. To accomplish this, the records of 225 flights are randomly selected and the number of unoccupied seats is noted for each of the sampled flights. The sample mean is 11.6 seats and the sample standard deviation is 4.1 seats.

a. i. x = ___________
   ii. s_x = ___________
   iii. n = ___________
   iv. n – 1 = ___________

b. Define the random variables X and X in words.
c. Which distribution should you use for this problem? Explain your choice.
d. Construct a 92% confidence interval for the population mean number of unoccupied seats per flight.

   i. State the confidence interval.
   ii. Calculate the error bound.

Solution:

a. i. x = 11.6
   ii. s_x = 4.1
   iii. n = 225
   iv. n – 1 = 224

b. X is the number of unoccupied seats on a single flight. X is the mean number of unoccupied seats from a sample of 225 flights.

c. We will use a Student's-t distribution because we do not know the population standard deviation.

d.
   i. CI: (11.12 , 12.08)
   ii. *EBM*: 0.48

Exercise:

Problem:

In a recent sample of 84 used car sales costs, the sample mean was $6,425 with a standard deviation of $3,156. Assume the underlying distribution is approximately normal.

a. Which distribution should you use for this problem? Explain your choice.
b. Define the random variable X in words.
c. Construct a 95% confidence interval for the population mean cost of a used car.

   i. State the confidence interval.
   ii. Calculate the error bound.

d. Explain what a "95% confidence interval" means for this study.

Exercise:

Problem:

Six different national brands of chocolate chip cookies were randomly selected at the supermarket. The grams of fat per serving are as follows: 8; 8; 10; 7; 9; 9. Assume the underlying distribution is approximately normal.

a. Construct a 90% confidence interval for the population mean grams of fat per serving of chocolate chip cookies sold in supermarkets.

   i. State the confidence interval.
   ii. Calculate the error bound.

b. If you wanted a smaller error bound while keeping the same level of confidence, what should have been changed in the study before it was done?
c. Go to the store and record the grams of fat per serving of six brands of chocolate chip cookies.
d. Calculate the mean.
e. Is the mean within the interval you calculated in part a? Did you expect it to be? Why or why not?

Solution:

a.
   i. CI: (7.64 , 9.36)

    ii. *EBM*: 0.86

b. The sample should have been increased.
c. Answers will vary.
d. Answers will vary.
e. Answers will vary.

Exercise:

Problem:

A survey of the mean number of cents off that coupons give was conducted by randomly surveying one coupon per page from the coupon sections of a recent San Jose Mercury News. The following data were collected: 20¢; 75¢; 50¢; 65¢; 30¢; 55¢; 40¢; 40¢; 30¢; 55¢; $1.50; 40¢; 65¢; 40¢. Assume the underlying distribution is approximately normal.

a.
    i. x = ___________
    ii. s_x = ___________
    iii. n = ___________
    iv. $n - 1$ = ___________

b. Define the random variables X and X in words.
c. Which distribution should you use for this problem? Explain your choice.
d. Construct a 95% confidence interval for the population mean worth of coupons.

    i. State the confidence interval.
    ii. Calculate the error bound.

e. If many random samples were taken of size 14, what percent of the confidence intervals constructed should contain the population mean worth of coupons? Explain why.

Use the following information to answer the next two exercises: A quality control specialist for a restaurant chain takes a random sample of size 12 to check the amount of soda served in the 16 oz. serving size. The sample mean is 13.30 with a sample standard deviation of 1.55. Assume the underlying population is normally distributed.

Exercise:

Problem:

Find the 95% Confidence Interval for the true population mean for the amount of soda served.

a. (12.42, 14.18)
b. (12.32, 14.29)
c. (12.50, 14.10)

d. Impossible to determine

Solution:

b

Exercise:

Problem: What is the error bound?

a. 0.87
b. 1.98
c. 0.99
d. 1.74

Glossary

Student's **t**-Distribution

investigated and reported by William S. Gossett in 1908 and published under the pseudonym Student; the major characteristics of the random variable (RV) are:

- It is continuous and assumes any real values.
- The pdf is symmetrical about its mean of zero. However, it is more spread out and flatter at the apex than the normal distribution.
- It approaches the standard normal distribution as n get larger.
- There is a "family of t–distributions: each representative of the family is completely defined by the number of degrees of freedom, which is one less than the number of data.

Estimating a Population Proportion

During an election year, we see articles in the newspaper that state **confidence intervals** in terms of proportions or percentages. For example, a poll for a particular candidate running for president might show that the candidate has 40% of the vote within three percentage points (if the sample is large enough). Often, election polls are calculated with 95% confidence, so the pollsters would be 95% confident that the true proportion of voters who favored the candidate would be between 0.37 and 0.43: (0.40 – 0.03, 0.40 + 0.03).

Investors in the stock market are interested in the true proportion of stocks that go up and down each week. Businesses that sell personal computers are interested in the proportion of households in the United States that own personal computers. Confidence intervals can be calculated for the true proportion of stocks that go up or down each week and for the true proportion of households in the United States that own personal computers.

The procedure to find the confidence interval, the sample size, the **error bound**, and the **confidence level** for a proportion is similar to that for the population mean, but the formulas are different.

How do you know you are dealing with a proportion problem? First, when you are dealing with a proportion problem, there will be no mention of a mean or average. Instead, the variable of interest is the proportion (often mentioned as a percent) of the population that falls into a particular category. If an individual falls into the category of interest, we call that a success. If X is the random variable that represents the number of successes in a sample and n is the total number of individuals in the sample (the sample size), then to form a sample proportion, take X, and divide it by n.

The random variable $\widehat{P}$ (read "P hat") is that proportion,

$$\widehat{P} = \frac{X}{n}$$

Similar to the sampling distribution of the sample mean, the **sampling distribution of the sample proportion** is the distribution of every sample proportion that can be calculated from every sample of the same sample size n, selected from the same population.

When n is large and p, the population proportion, is not close to zero or one, we can use the **normal distribution** to approximate the sampling distribution of $\widehat{p}$, where $\widehat{p}$ is the

sample proportion.

$\frac{X}{n} = \widehat{P} \sim N\left(p, \sqrt{\frac{pq}{n}}\right)$, where $q = 1 - p$.

The confidence interval has the form ($\widehat{p} - EBP$, $\widehat{p} + EBP$), where EBP is the **error bound for the proportion**.

$\widehat{p}$ = the **estimated proportion** of successes ($\widehat{p}$ is a **point estimate** for p, the true population proportion.)

$\widehat{p} = \frac{x}{n}$

x = the **number** of successes in the sample.

n = the size of the sample.

The error bound for a proportion is given by the following formula:

$EBP = \left(z_{\frac{\alpha}{2}}\right)\left(\sqrt{\frac{\widehat{p}\hat{q}}{n}}\right)$ where $\hat{q} = 1 - \widehat{p}$ and $z_{\frac{\alpha}{2}}$ is the z-score (also known as the critical value) associated with the confidence level.

This formula is similar to the error bound formula for a mean, except that the "appropriate standard deviation" is different. For a mean, when the population standard deviation is known, the appropriate standard deviation that we use is $\frac{\sigma}{\sqrt{n}}$. For a proportion, the appropriate standard deviation is $\sqrt{\frac{pq}{n}}$.

However, in the error bound formula, we use $\sqrt{\frac{\widehat{p}\hat{q}}{n}}$ to estimate the standard deviation of the sampling distribution of the sample proportion, $\sqrt{\frac{pq}{n}}$.

In the error bound formula, the **sample proportions $\widehat{p}$ and $\hat{q}$ are estimates of the unknown population proportions p and q**. The estimated proportions $\widehat{p}$ and $\hat{q}$ are used because p and q are not known. The sample proportions $\widehat{p}$ and $\hat{q}$ are calculated from the data: $\widehat{p}$ is the estimated proportion of successes, and $\hat{q}$ is the estimated proportion of failures.

Note: In general, the term **sample proportion** refers to the proportion of success in the sample, $\widehat{p}$.

The confidence interval can be used only if the number of successes $n\widehat{p}$ (or x) and the number of failures $n\hat{q}$ (or $n - x$) are both greater than ten.

Example:
Exercise:

Problem:

Suppose that a market research firm is hired to estimate the percent of adults living in a large city who have cell phones. Five hundred randomly selected adult residents in this city are surveyed to determine whether they have cell phones. Of the 500 people surveyed, 421 responded yes - they own cell phones. Using a 95% confidence level, compute a confidence interval estimate for the true proportion of adult residents of this city who have cell phones.

- The first solution is step-by-step (Solution A).
- The second solution uses a function of the TI-83, 83+ or 84 calculators (Solution B).

Solution:

To calculate the confidence interval, you must find $\widehat{p}$, $\hat{q}$, and *EBP*.

n = 500

Let x = the number of people in the sample who have cell phones.

x = the number of successes = 421. (Note that this number and 79, the number of failures, are both greater than 10.)

$\widehat{p} = \frac{x}{n} = \frac{421}{500} = 0.842$

$\widehat{p}$ = 0.842 is the sample proportion; this is the point estimate of the population proportion.

$\hat{q}$ = 1 – $\widehat{p}$ = 1 – 0.842 = 0.158

Since CL = 0.95, then α = 1 – CL = 1 – 0.95 = 0.05, which means $\left(\frac{\alpha}{2}\right)$ = 0.025.

Then $z_{\frac{\alpha}{2}} = z_{0.025} = 1.96$.

Use the TI-83, 83+, or 84+ calculator command invNorm(0.975,0,1) to find $z_{0.025}$. Remember that the area to the right of $z_{0.025}$ is 0.025 and the area to the left of $z_{0.025}$ is 0.975. This can also be found using appropriate commands on other calculators, using a computer, or using a Standard Normal probability table.

$$EBP = \left(z_{\frac{\alpha}{2}}\right)\sqrt{\frac{\hat{p}\hat{q}}{n}} = (1.96)\sqrt{\frac{(0.842)(0.158)}{500}} = 0.032$$

$$\hat{p} – EBP = 0.842 – 0.032 = 0.81$$

$$\hat{p} + EBP = 0.842 + 0.032 = 0.874$$

The confidence interval for the true population proportion is ($\hat{p}$ – EBP, $\hat{p}$ + EBP) = (0.810, 0.874).

Interpretation
We estimate with 95% confidence that between 81% and 87.4% of all adult residents of this city have cell phones.

Explanation of 95% Confidence Level
Ninety-five percent of the confidence intervals constructed in this way would contain the true value for the population proportion of all adult residents of this city who have cell phones.

Solution:

Solution B

Note:
Press `STAT` and arrow over to `TESTS`.
Arrow down to `A:1-PropZint`. Press `ENTER`.
Arrow down to x: and enter 421.
Arrow down to n: and enter 500.
Arrow down to `C-Level` and enter .95.
Arrow down to `Calculate` and press `ENTER`.
The confidence interval is (0.81003, 0.87397).

Note:
Try It
Exercise:

Problem:

Suppose 250 randomly selected people are surveyed to determine if they own a tablet. Of the 250 surveyed, 98 reported owning a tablet. Using a 95% confidence level, compute a confidence interval estimate for the true proportion of people who own tablets.

Solution:

(0.33148, 0.45252)

Example:
Exercise:

Problem:

For a class project, a political science student at a large university wants to estimate the percent of students who are registered voters. He surveys 500 students and finds that 300 are registered voters. Compute a 90% confidence interval for the true percent of students who are registered voters, and interpret the confidence interval.

- The first solution is step-by-step (Solution A).
- The second solution uses a function of the TI-83, 83+, or 84 calculators (Solution B).

Solution:

x = 300 and n = 500

$\widehat{p} = \frac{x}{n} = \frac{300}{500} = 0.6$

$\hat{q} = 1 - \widehat{p} = 1 - 0.6 = 0.4$

Since CL = 0.90, then α = 1 – CL = 1 – 0.90 = 0.10, which means $\left(\frac{\alpha}{2}\right)$ = 0.05

$z_{\frac{\alpha}{2}} = z_{0.05} = 1.645$

Use the TI-83, 83+, or 84+ calculator command invNorm(0.95,0,1) to find $z_{0.05}$. Remember that the area to the right of $z_{0.05}$ is 0.05 and the area to the left of $z_{0.05}$ is 0.95. This can also be found using appropriate commands on other calculators, using a computer, or using a standard normal probability table.

$$EBP = \left(z_{\frac{\alpha}{2}}\right)\sqrt{\frac{\hat{p}\hat{q}}{n}} = (1.645)\sqrt{\frac{(0.6)(0.4)}{500}} = 0.036$$

$$\hat{p} - EBP = 0.6 - 0.036 = 0.564$$

$$\hat{p} + EBP = 0.6 + 0.036 = 0.636$$

The confidence interval for the true population proportion is ($\hat{p} - EBP$, $\hat{p} + EBP$) = (0.564, 0.636).

Interpretation

- We estimate with 90% confidence that the true percent of all students that are registered voters is between 56.4% and 63.6%.
- Alternate Wording: We estimate with 90% confidence that between 56.4% and 63.6% of ALL students are registered voters.

Explanation of 90% Confidence Level
Ninety percent of all confidence intervals constructed in this way contain the true value for the population percent of students that are registered voters.

Solution:

Solution B

Note:
Press `STAT` and arrow over to `TESTS`.
Arrow down to `A:1-PropZint`. Press `ENTER`.
Arrow down to `x:` and enter 300.
Arrow down to `n:` and enter 500.
Arrow down to `C-Level` and enter 0.90.
Arrow down to `Calculate` and press `ENTER`.
The confidence interval is (0.56396, 0.63604).

Note:
Try It
Exercise:

Problem:

A student polls his school to see if students in the school district are for or against the new legislation regarding school uniforms. She surveys 600 students and finds that 480 are against the new legislation.

a. Compute a 90% confidence interval for the true percent of students who are against the new legislation and interpret the confidence interval.

Solution:

(0.77314, 0.82686); We estimate with 90% confidence that the true percent of all students in the district who are against the new legislation is between 77.3% and 82.7%.

Exercise:

Problem:

b. In a sample of 300 students, 68% said they own an iPod and a smart phone. Compute a 97% confidence interval for the true percent of students who own an iPod and a smartphone.

Solution:
Solution A

$\widehat{p} = 0.68$

$\hat{q} = 1 – \widehat{p} = 1 – 0.68 = 0.32$

Since CL = 0.97, we know α = 1 – 0.97 = 0.03 and $\frac{\alpha}{2}$ = 0.015.

The area to the left of $z_{0.015}$ is 0.015, and the area to the right of $z_{0.015}$ is 1 – 0.015 = 0.985.

Using the TI 83, 83+, or 84+ calculator function InvNorm(.985,0,1),

$z_{0.015} = 2.17$

$$EPB = \left(z_{\frac{\alpha}{2}}\right)\sqrt{\frac{\widehat{p}\hat{q}}{n}} = 2.17\sqrt{\frac{0.68(0.32)}{300}} \approx 0.05844$$

$\hat{p}$ – *EPB* = 0.68 – 0.05844 = 0.62156

$\hat{p}$ + *EPB* = 0.68 + 0.0584 = 0.73844

We are 97% confident that the true proportion of all students who own an iPod and a smart phone is between 0.62156 and 0.73844.

Solution:
Solution B

Note:
Press `STAT` and arrow over to `TESTS`.
Arrow down to `A:1-PropZint`. Press `ENTER`.
Arrow down to x`:` and enter 0.68*300.
Arrow down to n`:` and enter 300.
Arrow down to `C-Level` and enter 0.97.
Arrow down to `Calculate` and press `ENTER`.
The confidence interval is (0.62156, 0.73844).

“Plus Four” Confidence Interval for *p* (Optional)

There is a certain amount of error introduced into the process of calculating a confidence interval for a proportion. Because we do not know the true proportion for the population, we are forced to use point estimates to calculate the appropriate standard deviation of the sampling distribution. Studies have shown that the resulting estimation of the standard deviation can be flawed.

Fortunately, there is a simple adjustment that allows us to produce more accurate confidence intervals. We simply pretend that we have four additional observations. Two of these observations are successes and two are failures. The new sample size, then, is n + 4, and the new count of successes is x + 2.

Computer studies have demonstrated the effectiveness of this method. It should be used when the confidence level desired is at least 90% and the sample size is at least ten.

Example:
Exercise:

Problem:

A random sample of 25 statistics students was asked: “Have you smoked a cigarette in the past week?” Six students reported smoking within the past week. Use the “plus-four” method to find a 95% confidence interval for the true proportion of statistics students who smoke.

Solution:
Solution A

Six students out of 25 reported smoking within the past week, so x = 6 and n = 25. Because we are using the “plus-four” method, we will use x = 6 + 2 = 8 and n = 25 + 4 = 29.

$$\widehat{p}=\frac{x}{n}=\frac{8}{29}\approx 0.276$$

$\hat{q}=1–\widehat{p}=1–0.276=0.724$

Since CL = 0.95, we know α = 1 – 0.95 = 0.05 and $\frac{\alpha}{2}$ = 0.025.

$z_{0.025}=1.96$

$EPB=\left(z_{\frac{\alpha}{2}}\right)\sqrt{\frac{\widehat{p}\widehat{q}}{n}}=(1.96)\sqrt{\frac{0.276(0.724)}{29}}\approx 0.163$

$\widehat{p}$ – EPB = 0.276 – 0.163 = 0.113

$\widehat{p}$ + EPB = 0.276 + 0.163 = 0.439

We are 95% confident that the true proportion of all statistics students who smoke cigarettes is between 0.113 and 0.439.

Solution:

Note:
Press STAT and arrow over to TESTS.
Arrow down to A:1-PropZint. Press ENTER.

Note:
Reminder
Remember that the plus-four method assume an additional four trials: two successes and two failures. You do not need to change the process for calculating the confidence interval; simply update the values of x and n to reflect these additional trials.

Arrow down to x and enter 8.
Arrow down to n and enter 29.
Arrow down to C-Level and enter 0.95.
Arrow down to Calculate and press ENTER.
The confidence interval is (0.11319, 0.43853).

Note:
Try It
Exercise:

Problem:

Out of a random sample of 65 freshmen at State University, 31 students have declared a major. Use the “plus-four” method to find a 96% confidence interval for the true proportion of freshmen at State University who have declared a major.

Solution:
Solution A

Using “plus four,” we have x = 31 + 2 = 33 and n = 65 + 4 = 69.

$$\widehat{p} = \frac{33}{69} \approx 0.478$$

$$\hat{q} = 1 - \widehat{p} = 1 - 0.478 = 0.522$$

Since CL = 0.96, we know α = 1 – 0.96 = 0.04 and $\frac{\alpha}{2}$ = 0.02.

$$z_{0.02} = 2.054$$

$$EPB = \left(z_{\frac{\alpha}{2}}\right)\sqrt{\frac{\widehat{p}\hat{q}}{n}} = (2.054)\left(\sqrt{\frac{(0.478)(0.522)}{69}}\right) = 0.124$$

$\widehat{p}$ – *EPB* = 0.478 – 0.124 = 0.354

$\widehat{p}$ + *EPB* = 0.478 + 0.124 = 0.602

We are 96% confident that between 35.4% and 60.2% of all freshmen at State U have declared a major.

Solution:
Solution B

Note:
Press STAT and arrow over to TESTS.
Arrow down to A:1-PropZint. Press ENTER.
Arrow down to x and enter 33.
Arrow down to n and enter 69.
Arrow down to C-Level and enter 0.96.
Arrow down to Calculate and press ENTER.
The confidence interval is (0.35476, 0.60177).

Example:
Exercise:

Problem:

The Berkman Center for Internet & Society at Harvard recently conducted a study analyzing the privacy management habits of teen internet users. In a group of 50 teens, 13 reported having more than 500 friends on Facebook. Use the “plus four” method to find a 90% confidence interval for the true proportion of teens who would report having more than 500 Facebook friends.

Solution:

Using “plus-four,” we have x = 13 + 2 = 15 and n = 50 + 4 = 54.

$\widehat{p} = \frac{15}{54} \approx 0.278$

$\hat{q} = 1 – \widehat{p} = 1 - 0.241 = 0.722$

Since CL = 0.90, we know α = 1 – 0.90 = 0.10 and $\frac{\alpha}{2}$ = 0.05.

$z_{0.05} = 1.645$

$$EPB = (z_{\frac{\alpha}{2}})\left(\sqrt{\frac{\hat{p}\hat{q}}{n}}\right) = (1.645)\left(\sqrt{\frac{(0.278)(0.722)}{54}}\right) \approx 0.100$$

$\hat{p}$ – *EPB* = 0.278 – 0.100 = 0.178

$\hat{p}$ + *EPB* = 0.278 + 0.100 = 0.378

We are 90% confident that between 17.8% and 37.8% of all teens would report having more than 500 friends on Facebook.

Solution:

Note:

Press STAT and arrow over to TESTS.
Arrow down to A:1-PropZint. Press ENTER.
Arrow down to x and enter 15.
Arrow down to n and enter 54.
Arrow down to C-Level and enter 0.90.
Arrow down to Calculate and press ENTER.
The confidence interval is (0.17752, 0.37803).

Note:
Try It
Exercise:

Problem:

The Berkman Center Study referenced in [link] talked to teens in smaller focus groups, but also interviewed additional teens over the phone. When the study was complete, 588 teens had answered the question about their Facebook friends with 159 saying that they have more than 500 friends. Use the “plus-four” method to find a 90% confidence interval for the true proportion of teens that would report having more than 500 Facebook friends based on this larger sample. Compare the results to those in [link].

Solution:
Solution A

Using “plus-four,” we have x = 159 + 2 = 161 and n = 588 + 4 = 592.

$\widehat{p} = \frac{161}{592} \approx 0.272$

$\hat{q} = 1 - \widehat{p} = 1 - 0.272 = 0.728$

Since CL = 0.90, we know α = 1 – 0.90 = 0.10 and $\frac{\alpha}{2}$ = 0.05

$$EPB = \left(z_{\frac{\alpha}{2}}\right)\left(\sqrt{\frac{\widehat{p}\hat{q}}{n}}\right) = (1.645)\left(\sqrt{\frac{(0.272)(0.728)}{592}}\right) \approx 0.030$$

$\widehat{p}$ – EPB = 0.272 – 0.030 = 0.242

$\widehat{p}$ + EPB = 0.272 + 0.030 = 0.302

We are 90% confident that between 24.2% and 30.2% of all teens would report having more than 500 friends on Facebook.

Solution:
Solution B

Note:
Press STAT and arrow over to TESTS.
Arrow down to A:1-PropZint. Press ENTER.
Arrow down to x and enter 161.
Arrow down to n and enter 592.
Arrow down to C-Level and enter 0.90.
Arrow down to Calculate and press ENTER.
The confidence interval is (0.24188, 0.30204).

Conclusion: The confidence interval for the larger sample is narrower than the interval from [link]. Larger samples will always yield more precise confidence intervals than smaller samples. The “plus four” method has a greater impact on the smaller sample. It shifts the point estimate from 0.26 (13/50) to 0.278 (15/54). It has a smaller impact on the *EPB*, changing it from 0.102 to 0.100. In the larger sample, the point estimate undergoes a smaller shift: from 0.270 (159/588) to 0.272 (161/592). It is easy to see that the plus-four method has the greatest impact on smaller samples.

Calculating the Sample Size *n*

If researchers desire a specific margin of error, then they can use the error bound formula to calculate the required sample size.

The error bound formula for a population proportion is
Equation:

$$EBP=\left(z_{\frac{\alpha}{2}}\right)\left(\sqrt{\frac{\widehat{p}\hat{q}}{n}}\right)$$

Solving for n, we obtain a formula for determining the minimum sample size required to estimate the population proportion within the desired margin of error (*EBP*):
Equation:

$$n=\frac{\left(z_{\frac{\alpha}{2}}\right)^{2}\left(\widehat{p}\hat{q}\right)}{EBP^{2}}$$

Note: Since this formula gives us the **minimum** sample size required, and in reality the sample size must be a whole number, ALWAYS round decimal values UP to the next whole number, rather than using basic rounding rules.

Note: Usually we want to know the required sample size BEFORE we actually collect any data, which means we don't know the value of $\widehat{p}$. So, we must use an estimation of $\widehat{p}$ to use the formula. To do this, we can either use a sample proportion from a similar study or we can use 0.5 as a **conservative estimate** of $\widehat{p}$, since (0.5)(1 - 0.5) = 0.25, which is the largest product possible when multiplying a proportion times its complement (1 minus the proportion). (Try other products: (0.6)(0.4) = 0.24; (0.3)(0.7) = 0.21; (0.2)(0.8) = 0.16 and so on). The largest possible product gives us the largest n.

Example:
Exercise:

Problem:

Suppose a mobile phone company wants to determine the current percentage of customers aged 50+ who use text messaging on their cell phones. How many customers aged 50+ should the company survey in order to be 90% confident that the estimated (sample) proportion is within three percentage points of the true population proportion of customers aged 50+ who use text messaging on their cell phones.

Solution:

From the problem, we know that ***EBP* = 0.03** (3% = 0.03) and $z_{\frac{\alpha}{2}}$ = $z_{0.05}$ = 1.645 because the confidence level is 90%.

However, in order to find n, we need to know the estimated (sample) proportion $\widehat{p}$. Remember that $\hat{q}$ = 1 – $\widehat{p}$. But, we do not know $\widehat{p}$ yet. Since we multiply $\widehat{p}$ and $\hat{q}$ together, we estimate them both to be equal to 0.5 because $\widehat{p}\hat{q}$ = (0.5)(0.5) = 0.25 results in the largest possible product. This gives us a large enough sample so that we can be 90% confident that we are within three percentage points of the true population proportion. To calculate the sample size n, use the formula and make the substitutions.

$n = \frac{z^2\widehat{p}\hat{q}}{EBP^2}$ gives $n = \frac{1.645^2(0.5)(0.5)}{0.03^2} = 751.7$

Round the answer UP to the next whole number. The sample size should be 752 cell phone customers aged 50+ in order to be 90% confident that the estimated (sample) proportion is within three percentage points of the true population proportion of all customers aged 50+ who use text messaging on their cell phones.

Note:
Try It
Exercise:

Problem:

Suppose an internet marketing company wants to determine the current percentage of customers who click on ads on their smartphones. How many customers should the company survey in order to be 95% confident that the estimated proportion is within five percentage points of the true population proportion of customers who click on ads on their smartphones?

Solution:

385 customers should be surveyed.

References

Jensen, Tom. “Democrats, Republicans Divided on Opinion of Music Icons.” Public Policy Polling. Available online at http://www.publicpolicypolling.com/Day2MusicPoll.pdf (accessed July 2, 2013).

Madden, Mary, Amanda Lenhart, Sandra Coresi, Urs Gasser, Maeve Duggan, Aaron Smith, and Meredith Beaton. “Teens, Social Media, and Privacy.” PewInternet, 2013. Available online at http://www.pewinternet.org/Reports/2013/Teens-Social-Media-And-Privacy.aspx (accessed July 2, 2013).

Prince Survey Research Associates International. “2013 Teen and Privacy Management Survey.” Pew Research Center: Internet and American Life Project. Available online at http://www.pewinternet.org/~/media//Files/Questionnaire/2013/Methods%20and%20Questions_Teens%20and%20Social%20Media.pdf (accessed July 2, 2013).

Saad, Lydia. “Three in Four U.S. Workers Plan to Work Pas Retirement Age: Slightly more say they will do this by choice rather than necessity.” Gallup® Economy, 2013. Available online at http://www.gallup.com/poll/162758/three-four-workers-plan-work-past-retirement-age.aspx (accessed July 2, 2013).

The Field Poll. Available online at http://field.com/fieldpollonline/subscribers/ (accessed July 2, 2013).

Zogby. “New SUNYIT/Zogby Analytics Poll: Few Americans Worry about Emergency Situations Occurring in Their Community; Only one in three have an Emergency Plan; 70% Support Infrastructure ‘Investment’ for National Security.” Zogby Analytics, 2013. Available online at http://www.zogbyanalytics.com/news/299-americans-neither-worried-nor-prepared-in-case-of-a-disaster-sunyit-zogby-analytics-poll (accessed July 2, 2013).

"52% Say Big-Time College Athletics Corrupt Education Process." Rasmussen Reports, 2013. Available online at http://www.rasmussenreports.com/public_content/lifestyle/sports/may_2013/52_say_big_time_college_athletics_corrupt_education_process (accessed July 2, 2013).

Section Review

Some statistical measures, like many survey questions, measure qualitative rather than quantitative data. In this case, the population parameter being estimated is a proportion. It is possible to create a confidence interval for the true population proportion following procedures similar to those used in creating confidence intervals for population means. The formulas are slightly different, but they follow the same reasoning.

Let $\widehat{p}$ represent the sample proportion, x/n, where x represents the number of successes and n represents the sample size. Let $\hat{q}$ = 1 – $\widehat{p}$. Then the confidence interval for a population proportion is given by the following formula:

(lower bound, upper bound) $= (\widehat{p} - EBP, \widehat{p} + EBP) = \left(\widehat{p} - z_{\frac{\alpha}{2}}\sqrt{\frac{\widehat{p}\hat{q}}{n}}, \widehat{p} + z_{\frac{\alpha}{2}}\sqrt{\frac{\widehat{p}\hat{q}}{n}}\right)$, where $\hat{q}$ = 1–$\widehat{p}$ and $z_{\frac{\alpha}{2}}$ is the z-score associated with the confidence level.

The "plus four" method for calculating confidence intervals is an attempt to balance the error introduced by using estimates of the population proportion when calculating the standard deviation of the sampling distribution. Simply imagine four additional trials in the study; two are successes and two are failures. Calculate $\widehat{p} = \frac{x+2}{n+4}$, and proceed to find the confidence interval. When sample sizes are small, this method has been demonstrated to provide more accurate confidence intervals than the standard formula used for larger samples.

Formula Review

The sample proportion $\widehat{p}$ = x/n, where x represents the number of successes in the sample and n represents the sample size. The variable $\widehat{p}$ serves as the point estimate for the true population proportion.

$\hat{q}$ = 1 – $\widehat{p}$

The sampling distribution of $\widehat{p}$ can be approximated with the normal distribution shown here.

$\widehat{P} \sim N\left(p, \sqrt{\frac{pq}{n}}\right)$

EBP = the error bound for a proportion = $z_{\frac{\alpha}{2}}\sqrt{\frac{\widehat{p}\hat{q}}{n}}$

Confidence interval for a proportion:

$$(\text{lower bound, upper bound}) = (\widehat{p} - EBP, \widehat{p} + EBP) = \left(\widehat{p} - z_{\frac{\alpha}{2}}\sqrt{\frac{\widehat{p}\hat{q}}{n}},\ \widehat{p} + z_{\frac{\alpha}{2}}\sqrt{\frac{\widehat{p}\hat{q}}{n}}\right)$$

$$n = \frac{\left(z_{\frac{\alpha}{2}}\right)^2\left(\widehat{p}\hat{q}\right)}{EBP^2}$$

provides the number of participants needed to estimate the population proportion with confidence 1 - α and margin of error *EBP*.

Use the following information to answer the next two exercises: Marketing companies are interested in knowing the population percent of women who make the majority of household purchasing decisions.

Exercise:

Problem:

When designing a study to determine this population proportion, what is the minimum number you would need to survey to be 90% confident that the population proportion is estimated to within 0.05?

Exercise:

Problem:

If it were later determined that it was important to be more than 90% confident and a new survey were commissioned, how would it affect the minimum number you need to survey? Why?

Solution:

It would increase, because the z-score would increase, which increases the numerator and thereby increases the number.

Use the following information to answer the next five exercises: Suppose the marketing company did a survey. They randomly surveyed 200 households and found that in 120 of them, the woman made the majority of the purchasing decisions. We are interested in the proportion of households where women make the majority of the purchasing decisions.

Exercise:

Problem: Identify the following:

a. x = _______
b. n = _______
c. $\widehat{p}$ = _______

Exercise:

Problem: Define the random variables X and $\widehat{P}$ in words.

Solution:

X is the number of “successes”, in the households sampled, where the woman makes the majority of the purchasing decisions for the household. $\widehat{P}$ is the percentage of households sampled where the woman makes the majority of the purchasing decisions for the household.

Exercise:

Problem: Which distribution should you use for this problem?

Exercise:

Problem:

Construct a 95% confidence interval for the population proportion of households where the women make the majority of the purchasing decisions. State the confidence interval and calculate the error bound.

Solution:

CI: (0.5321, 0.6679)

EBM: 0.0679

Exercise:

Problem:

List two difficulties the company might have in obtaining random results, if this survey were done by email.

Use the following information to answer the next five exercises: Of 1,050 randomly selected adults, 360 identified themselves as manual laborers, 280 identified themselves as non-manual wage earners, 250 identified themselves as mid-level managers, and 160 identified themselves as executives. In the survey, 82% of manual laborers preferred trucks, 62% of non-manual wage earners preferred trucks, 54% of mid-level managers preferred trucks, and 26% of executives preferred trucks.

Exercise:

Problem:

We are interested in finding the 95% confidence interval for the percent of executives who prefer trucks. Define random variables X and $\widehat{P}$ in words.

Solution:

X is the number of “successes”, in the sample, where an executive prefers a truck. $\widehat{P}$ is the percentage of executives sampled who prefer a truck.

Exercise:

Problem: Estimate the distribution which you should use for this problem.

Exercise:

Problem:

Construct a 95% confidence interval. State the confidence interval and calculate the error bound.

Solution:

CI: (0.19432, 0.33068)

EBM = 0.0707

Exercise:

Problem:

Suppose we want to lower the sampling error. What is one way to accomplish that?

Exercise:

Problem:

Suppose the company increased the confidence level, but used the same data. What effect would this have on the interval?

Solution:

Increasing the confidence level would increase the z-score, which would increase the error bound. This in turn would widen the interval.

Use the following information to answer the next five exercises: A poll of 1,200 voters asked what the most significant issue was in the upcoming election. Sixty-five percent answered the economy. We are interested in the proportion of voters who feel the economy is the most important.

Exercise:

Problem: Define the random variable X in words.

Exercise:

Problem: Define the random variable $\widehat{P}$ in words.

Solution:

$\widehat{P}$ is the proportion of voters sampled who said the economy is the most important issue in the upcoming election.

Exercise:

Problem: Which distribution should you use for this problem?

Exercise:

Problem:

Construct a 90% confidence interval and state the confidence interval and the error bound.

Solution:

CI: (0.62735, 0.67265)

EBM = 0.02265

Exercise:

Problem:

What would happen to the confidence interval if the level of confidence were 95%?

Use the following information to answer the next 16 exercises: The Ice Chalet offers dozens of different beginning ice-skating classes. All of the class names are put into a bucket. The 5 P.M., Monday night, ages 8 to 12, beginning ice-skating class was picked. In that class were 64 girls and 16 boys. Suppose that we are interested in the true proportion of girls, ages 8 to 12, in all beginning ice-skating classes at the Ice Chalet. Assume that the children in the selected class are a random sample of the population.

Exercise:

Problem: What is being counted?

Solution:

The number of girls, ages 8 to 12, in the 5 P.M. Monday night beginning ice-skating class.

Exercise:

Problem: In words, define the random variable X.

Exercise:

Problem: Calculate the following:

a. x = ________
b. n = ________
c. $\widehat{p}$ = ________

Solution:

a. x = 64
b. n = 80
c. $\widehat{p}$ = 0.8

Exercise:

Problem: Define a new random variable $\widehat{P}$. What is $\widehat{p}$ estimating?

Solution:

p

Exercise:

Problem: In words, define the random variable $\widehat{P}$.

Exercise:

Problem:

State the estimated distribution of $\widehat{P}$. Construct a 92% Confidence Interval for the true proportion of girls in the ages 8 to 12 beginning ice-skating classes at the Ice Chalet.

Solution:

$\widehat{P} \sim N\left(0.8, \sqrt{\frac{(0.8)(0.2)}{80}}\right)$. CI: (0.72171, 0.87829).

Exercise:

Problem:

When looking at the distribution to determine the **z-**score, how much area is in both tails (combined)?

Exercise:

Problem: How much area is in each tail?

Solution:

0.04

Exercise:

Problem: Calculate the following:

a. lower limit
b. upper limit
c. error bound

Exercise:

Problem: The 92% confidence interval is _______. Round to two decimal places.

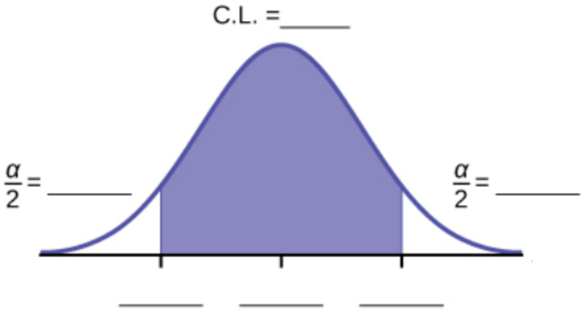
Solution:

(0.72, 0.88)

Exercise:

Problem:

Fill in the blanks on the graph with the areas and upper and lower **z**-scores associated with the confidence level, as well as the mean of this distribution.



Exercise:

Problem: In one complete sentence, explain what the interval means.

Solution:

With 92% confidence, we estimate the proportion of girls, ages 8 to 12, in a beginning ice-skating class at the Ice Chalet to be between 72% and 88%.

Exercise:

Problem:

Using the same $\hat{p}$ and level of confidence, suppose that n were increased to 100. Would the error bound become larger or smaller? How do you know?

Exercise:

Problem:

Using the same $\hat{p}$ and n = 80, how would the error bound change if the confidence level were increased to 98%? Why?

Solution:

The error bound would increase. Assuming all other variables are kept constant, as the confidence level increases, the area under the curve corresponding to the confidence level becomes larger, which creates a wider interval and thus a larger error.

Exercise:

Problem:

If you decreased the allowable error bound, why would the minimum sample size increase (keeping the same level of confidence)?

Homework

Exercise:

Problem:

Insurance companies are interested in knowing the population percent of drivers who always buckle up before riding in a car.

a. When designing a study to determine this population proportion, what is the minimum number you would need to survey to be 95% confident that the population proportion is estimated to within 0.03?
b. If it were later determined that it was important to be more than 95% confident and a new survey was commissioned, how would that affect the minimum number you would need to survey? Why?

Solution:

a. 1,068
b. The sample size would need to be increased since the critical value increases as the confidence level increases.

Exercise:

Problem:

Suppose that the insurance companies did a survey. They randomly surveyed 400 drivers and found that 320 claimed they always buckle up. We are interested in the true proportion of drivers who claim they always buckle up.

a.
    i. x = ___________
    ii. n = ___________
    iii. $\widehat{p}$ = ___________

b. Define the random variables X and $\widehat{P}$, in words.
c. Which distribution should you use for this problem? Explain your choice.
d. Construct a 95% confidence interval for the population proportion who claim they always buckle up.

   i. State the confidence interval.
   ii. Calculate the error bound.

e. If this survey were done by telephone, list three difficulties the companies might have in obtaining random results.

Exercise:

Problem:

According to a recent survey of 1,200 people, 61% feel that the president is doing an acceptable job. We are interested in the population proportion of people who feel the president is doing an acceptable job.

a. Define the random variables X and $\widehat{P}$ in words.
b. Which distribution should you use for this problem? Explain your choice.
c. Construct a 90% confidence interval for the population proportion of people who feel the president is doing an acceptable job.

   i. State the confidence interval.
   ii. Calculate the error bound.

Solution:

a. X = the number of people in the sample who feel that the president is doing an acceptable job;

   $\widehat{P}$ = the proportion of people in a sample who feel that the president is doing an acceptable job.
b. Since we're estimating a single population proportion and the number of successes and failures in the sample are both more than 10, given $\widehat{p}$ = 0.61 and n = 1200, we should use the following distribution: $N\left(0.61, \sqrt{\frac{(0.61)(0.39)}{1200}}\right)$

c. i. CI: (0.59, 0.63)
   ii. *EBM* = 0.02

Exercise:

Problem:

An article regarding interracial dating and marriage recently appeared in the Washington Post. Of the 1,709 randomly selected adults, 315 identified themselves as Latinos, 323 identified themselves as blacks, 254 identified themselves as Asians, and 779 identified themselves as whites. In this survey, 86% of blacks said that they would welcome a white person into their families. Among Asians, 77% would welcome a white person into their families, 71% would welcome a Latino, and 66% would welcome a black person.

a. We are interested in finding the 95% confidence interval for the percent of all black adults who would welcome a white person into their families. Define the random variables X and $\widehat{P}$, in words.
b. Which distribution should you use for this problem? Explain your choice.
c. Construct a 95% confidence interval.
   i. State the confidence interval.
   ii. Calculate the error bound.

Exercise:

Problem: Refer to the information in the previous exercise.

a. Construct three 95% confidence intervals.
   i. percent of all Asians who would welcome a white person into their families.
   ii. percent of all Asians who would welcome a Latino into their families.
   iii. percent of all Asians who would welcome a black person into their families.
b. Even though the three point estimates are different, do any of the confidence intervals overlap? Which?
c. For any intervals that do overlap, in words, what does this imply about the significance of the differences in the true proportions?
d. For any intervals that do not overlap, in words, what does this imply about the significance of the differences in the true proportions?

Solution:

a.
   i. (0.72, 0.82)
   ii. (0.65, 0.76)

    iii. (0.60, 0.72)

b. Yes, the intervals (0.72, 0.82) and (0.65, 0.76) overlap, and the intervals (0.65, 0.76) and (0.60, 0.72) overlap.
c. We can say that there does not appear to be a significant difference between the proportion of Asian adults who say that their families would welcome a white person into their families and the proportion of Asian adults who say that their families would welcome a Latino person into their families.
d. We can say that there is a significant difference between the proportion of Asian adults who say that their families would welcome a white person into their families and the proportion of Asian adults who say that their families would welcome a black person into their families.

Exercise:

Problem:

Stanford University conducted a study of whether running is healthy for men and women over age 50. During the first eight years of the study, 1.5% of the 1451 members of the 50-Plus Fitness Association died. We are interested in the proportion of people over 50 who ran and died in the same eight-year period.

a. Define the random variables X and $\widehat{P}$ in words.
b. Which distribution should you use for this problem? Explain your choice.
c. Construct a 97% confidence interval for the population proportion of people over 50 who ran and died in the same eight–year period.

    i. State the confidence interval.
    ii. Calculate the error bound.

d. Explain what a “97% confidence interval” means for this study.

Exercise:

Problem:

A telephone poll of 1,000 adult Americans was reported in an issue of Time Magazine. One of the questions asked was “What is the main problem facing the country?” Twenty percent answered “crime.” We are interested in the population proportion of adult Americans who feel that crime is the main problem.

a. Define the random variables X and $\widehat{P}$ in words.
b. Which distribution should you use for this problem? Explain your choice.
c. Construct a 95% confidence interval for the population proportion of adult Americans who feel that crime is the main problem.

    i. State the confidence interval.
    ii. Calculate the error bound.

d. Suppose we want to lower the sampling error. What is one way to accomplish that?
e. The sampling error given by Yankelovich Partners, Inc. (which conducted the poll) is ±3%. In one to three complete sentences, explain what the ±3% represents.

Solution:

a. X = the number of adult Americans who feel that crime is the main problem; $\hat{P}$ = the proportion of adult Americans who feel that crime is the main problem
b. Since we are estimating a proportion and the number of successes and failures in the sample are both more than 10, given $\hat{p}$ = 0.2 and n = 1000, the distribution we should use is $N\left(0.2, \sqrt{\frac{(0.2)(0.8)}{1000}}\right)$.

c.
    i. CI: (0.18, 0.22)
    ii. *EBM* = 0.02

d. One way to lower the sampling error is to increase the sample size.
e. The stated "± 3%" represents the maximum error bound. This means that those doing the study are reporting a maximum error of 3%.

Exercise:

Problem:

Refer to the previous exercise. Another question in the poll was "[How much are] you worried about the quality of education in our schools?" Sixty-three percent responded "a lot". We are interested in the population proportion of adult Americans who are worried a lot about the quality of education in our schools.

a. Define the random variables X and $\hat{P}$ in words.
b. Which distribution should you use for this problem? Explain your choice.
c. Construct a 95% confidence interval for the population proportion of adult Americans who are worried a lot about the quality of education in our schools.

    i. State the confidence interval.
    ii. Calculate the error bound.

d. The sampling error given by Yankelovich Partners, Inc. (which conducted the poll) is ±3%. In one to three complete sentences, explain what the ±3% represents.

Use the following information to answer the next three exercises: According to a Field Poll, 79% of California adults (actual results are 400 out of 506 surveyed) feel that “education and our schools” is one of the top issues facing California. We wish to construct a 90% confidence interval for the true proportion of California adults who feel that education and the schools is one of the top issues facing California.

Exercise:

Problem: A point estimate for the true population proportion is:

a. 0.90
b. 1.27
c. 0.79
d. 400

Solution:

c

Exercise:

Problem: A 90% confidence interval for the population proportion is _______.

a. (0.761, 0.820)
b. (0.125, 0.188)
c. (0.755, 0.826)
d. (0.130, 0.183)

Exercise:

Problem: The error bound is approximately _____.

a. 1.581
b. 0.791
c. 0.059
d. 0.030

Solution:

d

Use the following information to answer the next two exercises: Five hundred and eleven (511) homes in a certain southern California community are randomly surveyed to determine if they meet minimal earthquake preparedness recommendations. One hundred seventy-three (173) of the homes surveyed met the minimum recommendations for earthquake preparedness, and 338 did not.

Exercise:

Problem:

Find the confidence interval at the 90% Confidence Level for the true population proportion of southern California community homes meeting at least the minimum recommendations for earthquake preparedness.

a. (0.2975, 0.3796)
b. (0.6270, 0.6959)
c. (0.3041, 0.3730)
d. (0.6204, 0.7025)

Exercise:

Problem:

The point estimate for the population proportion of homes that do not meet the minimum recommendations for earthquake preparedness is ______.

a. 0.6614
b. 0.3386
c. 173
d. 338

Solution:

a

Exercise:

Problem:

On May 23, 2013, Gallup reported that of the 1,005 people surveyed, 76% of U.S. workers believe that they will continue working past retirement age. The confidence level for this study was reported at 95% with a ±3% margin of error.

a. Determine the estimated proportion from the sample.
b. Determine the sample size.
c. Identify CL and α.
d. Calculate the error bound based on the information provided.
e. Compare the error bound in part d to the margin of error reported by Gallup. Explain any differences between the values.
f. Create a confidence interval for the results of this study.
g. A reporter is covering the release of this study for a local news station. How should she explain the confidence interval to her audience?

Exercise:

Problem:

A national survey of 1,000 adults was conducted on May 13, 2013 by Rasmussen Reports. It concluded with 95% confidence that 49% to 55% of Americans believe that big-time college sports programs corrupt the process of higher education.

a. Find the point estimate and the error bound for this confidence interval.
b. Can we (with 95% confidence) conclude that more than half of all American adults believe this?
c. Use the point estimate from part a and n = 1,000 to calculate a 75% confidence interval for the proportion of American adults that believe that major college sports programs corrupt higher education.
d. Can we (with 75% confidence) conclude that at least half of all American adults believe this?

Solution:

a. $\widehat{p}$ = $\frac{(0.55+0.49)}{2}$ = 0.52; *EBP* = 0.55 - 0.52 = 0.03
b. No, the confidence interval includes values less than or equal to 0.50. It is possible that less than half of the population believe this.
c. CL = 0.75, so α = 1 – 0.75 = 0.25 and $\frac{\alpha}{2} = 0.125\ z_{\frac{\alpha}{2}} = 1.150$. (The area to the right of this z is 0.125, so the area to the left is 1 – 0.125 = 0.875.)

$EBP = (1.150)\sqrt{\frac{0.52(0.48)}{1{,}000}} \approx 0.018$
($\widehat{p}$ - *EBP*, $\widehat{p}$ + *EBP*) = (0.52 – 0.018, 0.52 + 0.018) = (0.502, 0.538)

Alternate Solution:

STAT TESTS A: 1-PropZinterval with x = (0.52)(1,000), n = 1,000, CL = 0.75.

Answer is (0.50183, 0.53817)

d. Yes – this interval does not fall less than 0.50 so we can conclude that at least half of all American adults believe that major sports programs corrupt education – but we do so with only 75% confidence.

Exercise:

Problem:

Public Policy Polling recently conducted a survey asking adults across the U.S. about music preferences. When asked, 80 of the 571 participants admitted that they have illegally downloaded music.

a. Create a 99% confidence interval for the true proportion of American adults who have illegally downloaded music.
b. This survey was conducted through automated telephone interviews on May 6 and 7, 2013. The error bound of the survey compensates for sampling error, or natural variability among samples. List some factors that could affect the survey's outcome that are not covered by the margin of error.
c. Without performing any calculations, describe how the confidence interval would change if the confidence level changed from 99% to 90%.

Exercise:

Problem:

You plan to conduct a survey on your college campus to learn about the political awareness of students. You want to estimate the true proportion of college students on your campus who voted in the 2012 presidential election with 95% confidence and a margin of error no greater than five percent. How many students must you interview?

Solution:

CL = 0.95; α = 1 – 0.95 = 0.05; $\frac{\alpha}{2}$ = 0.025; $z_{\frac{\alpha}{2}}$ = 1.96. Use $\widehat{p}$ = $\hat{q}$ = 0.5.

$$n = \frac{\left(z_{\frac{\alpha}{2}}\right)^2\left(\hat{p}\hat{q}\right)}{EBP^2} = \frac{1.96^2(0.5)(0.5)}{0.05^2} = 384.16$$

You need to interview at least 385 students to estimate the proportion to within 5% at 95% confidence.

Exercise:

Problem:

In a recent Zogby International Poll, nine of 48 respondents rated the likelihood of a terrorist attack in their community as “likely” or “very likely.” Use the “plus four” method to create a 97% confidence interval for the proportion of American adults who believe that a terrorist attack in their community is likely or very likely. Explain what this confidence interval means in the context of the problem.

Glossary

Error Bound for a Population Proportion (EBP)
: the margin of error; depends on the confidence level, the sample size, and the estimated (from the sample) proportion of successes.

Sampling distribution of the sample proportion
: the distribution of every possible sample proportion that can be calculated from a sample of sample size n, selected from a population.

Hypothesis Testing With One Sample: Introduction class="introduction"

You can use a hypothesis test to decide if a dog breeder's claim that every Dalmatian has 35 spots is statistically sound. (Credit: Robert Neff)

Note:

Chapter Objectives

By the end of this chapter, the student should be able to:

- Describe hypothesis testing in general and in practice
- Differentiate between Type I and Type II Errors
- Conduct and interpret hypothesis tests for a single population mean, population standard deviation known.
- Conduct and interpret hypothesis tests for a single population mean, population standard deviation unknown.
- Conduct and interpret hypothesis tests for a single population proportion.

One job of a statistician is to make statistical inferences about populations based on samples taken from the population. **Confidence intervals** are one way to estimate a population parameter. Another way to make a statistical inference is to make a decision about a parameter. For instance, a car dealer advertises that its new small truck gets 35 miles per gallon, on average. A tutoring service claims that its method of tutoring helps 90% of its students get an A or a B. A company says that women managers in their company earn an average of $60,000 per year.

A statistician will make a decision about these claims. This process is called "**hypothesis testing**." A hypothesis test involves collecting data from a sample and evaluating the data. Then, the statistician makes a decision as to whether or not there is sufficient evidence, based upon analyses of the data, to reject the null hypothesis.

In this chapter, you will conduct hypothesis tests on single means and single proportions. You will also learn about the errors associated with these tests.

Hypothesis testing consists of two contradictory hypotheses or statements, a decision based on the data, and a conclusion. To perform a hypothesis test, a statistician will:

1. Set up two contradictory hypotheses.
2. Collect sample data (in homework problems, the data or summary statistics will be given to you).
3. Determine the correct distribution to perform the hypothesis test.
4. Analyze sample data by performing the calculations that ultimately will allow you to reject or decline to reject the null hypothesis.
5. Make a decision and write a meaningful conclusion.

Note:
Note

To do the hypothesis test homework problems for this chapter and later chapters, make copies of the appropriate special solution sheets. See Appendix A.

Glossary

Confidence Interval (CI)

an interval estimate for an unknown population parameter. This depends on:

- The desired confidence level.
- Information that is known about the distribution (for example, known standard deviation).
- The sample and its size.

Hypothesis Testing

Based on sample evidence, a procedure for determining whether the hypothesis stated is a reasonable statement and should not be rejected, or is unreasonable and should be rejected.

The Null and Alternative Hypotheses

The actual test begins by considering two **hypotheses**. They are called the **null hypothesis** and the **alternative hypothesis**.

H_0 : **The null hypothesis:** It is a statement about the population that either is believed to be true or is used to put forth an argument unless it can be shown to be incorrect beyond a reasonable doubt.

H_a : **The alternative hypothesis:** It is a claim about the population that we are looking for evidence to support and what we conclude when we reject H_0 .

When performing a hypothesis test, you must examine evidence to decide if you have enough evidence to reject the null hypothesis or not. The evidence is in the form of sample data.

After careful analysis of the sample data, you make a **decision.** There are two options for a decision. They are "reject H_0" if the sample information favors the alternative hypothesis or "fail to reject H_0" if the sample information is insufficient to reject the null hypothesis.

Mathematical Symbols Used in H_0 and H_a :

H_0	H_a
equal (=)	not equal (≠)

H_0	H_a
equal (=)	less than (<)
equal (=)	more than (>)

Often the null hypothesis is known as the hypothesis that there is no difference (or no effect) and the alternative hypothesis is known as the hypothesis that there is some difference (or effect), which may be directional or not.

Note:

Note

Notice that H_0 always has equal sign in it, while the H_a never has a symbol with an equal in it. The choice of symbol for H_a depends on the wording of the problem for which you are running a hypothesis test. Just remember that **the alternative hypothesis is what you are looking for evidence to support**.

Be aware that some researchers may use ≥ or ≤ in the null hypothesis, opposite > and < respectively as the symbol in the alternative hypothesis, rather than an equal sign. However, in either case, the alternative hypothesis will never include a symbol with an equal in it.

Example:

H_0 : Thiry percent of the registered voters in Santa Clara County voted in the primary election. $p = 0.30$

H_a : More than thirty percent of the registered voters in Santa Clara County voted in the primary election. $p > 0.30$

Note:
Try It
Exercise:

Problem:

A medical trial is conducted to test whether or not a new medicine reduces cholesterol by 25%. State the null and alternative hypotheses.

Solution:

H_0 : The drug reduces cholesterol by 25%. $p = 0.25$

H_a : The drug does not reduce cholesterol by 25%. $p \neq 0.25$

Example:
We want to test whether the mean GPA of students in American colleges is different from 2.0 (out of 4.0).

When writing the hypotheses, ask yourself "What am I looking for evidence of?" The answer to this question will help you write the alternative hypothesis. The null hypothesis will be the same, except with an equal sign.

In this example, we are looking for evidence that the **mean** GPA of students in American colleges is **different from** 2.0. Since the variable of interest here is a mean, we need to use μ and the phrase "different from" is our clue to use a ≠ sign in our alternative hypothesis.

Therefore, the null and alternative hypotheses are:

$H_0 : \mu = 2.0$
$H_a : \mu \neq 2.0$

Note:
Try It
Exercise:

Problem:

We want to test whether the mean height of eighth graders is 66 inches. State the null and alternative hypotheses. Fill in the correct symbol (=, ≠, <, >) for the null and alternative hypotheses.

a. $H_0 : \mu$ __ 66
b. $H_a : \mu$ __ 66

Solution:

a. $H_0 : \mu = 66$
b. $H_a : \mu \neq 66$

Example:
We want to test if college students take less than five years to graduate from college, on average.

What are we looking for evidence of?

We are looking for evidence that college students take **less than** 5 years, on **average** to graduate from college.

The null and alternative hypotheses are:

$H_0 : \mu = 5$
$H_a : \mu < 5$

Note:
Try It
Exercise:

Problem:

We want to test if it takes fewer than 45 minutes to teach a lesson plan. State the null and alternative hypotheses. Fill in the correct symbol (=, ≠, <, >) for the null and alternative hypotheses.

a. $H_0 : \mu$ __ 45
b. $H_a : \mu$ __ 45

Solution:

a. $H_0 : \mu = 45$
b. $H_a : \mu < 45$

Example:
In an issue of *U. S. News and World Report*, an article on school standards stated that about half of all students in France, Germany, and Israel take advanced placement exams and a third pass. The same article stated that 6.6% of U.S. students take advanced placement exams and 4.4% pass. Test if the percentage of U.S. students who take advanced placement exams is more than 6.6%. State the null and alternative hypotheses.

What are we looking for evidence of?

We are looking for evidence that the **percentage** of U.S. students who take advanced placement exams is **more than** 6.6%.

Therefore, the null and alternative hypotheses are:

$H_0 : p = 0.066$

$H_a : p > 0.066$

Remember that when the variable of interest is a proportion, which is usually expressed as a percent in the problem, a p is used to represent the variable and the percent must be written in decimal form within the hypotheses.

Note:
Try It
Exercise:

Problem:

On a state driver's test, about 40% pass the test on the first try. We want to test if more than 40% pass on the first try. Fill in the correct symbol (=, ≠, <, >) for the null and alternative hypotheses.

a. $H_0 : p$ __ 0.40
b. $H_a : p$ __ 0.40

Solution:

a. $H_0 : p = 0.40$
b. $H_a : p > 0.40$

Note:
Collaborative Exercise
Bring to class a newspaper, some news magazines, and some Internet articles . In groups, find articles from which your group can write null and alternative hypotheses. Discuss your hypotheses with the rest of the class.

Chapter Review

In a **hypothesis test**, sample data is evaluated in order to arrive at a decision about some type of claim. If certain conditions about the sample are satisfied, then the claim can be evaluated for a population. In a hypothesis test, we:

1. Evaluate the **null hypothesis**, typically denoted with H_0 . The null is not rejected unless the hypothesis test shows otherwise. The null statement must always contain the equality symbol (=).
2. Always write the **alternative hypothesis**, typically denoted with H_a , using less than, greater than, or not equals symbols, i.e., (≠, >, or <).
3. If we reject the null hypothesis, then we can assume there is enough evidence to support the alternative hypothesis.
4. Never state that a claim is proven true or false. Keep in mind the underlying fact that hypothesis testing is based on probability laws; therefore, we can talk only in terms of non-absolute certainties.

Practice

Exercise:

Problem:

You are testing that the mean speed of your cable Internet connection is more than three Megabits per second. What is the random variable? Describe in words.

Solution:

The random variable is the mean Internet speed in Megabits per second.

Exercise:

Problem:

You are testing that the mean speed of your cable Internet connection is more than three Megabits per second. State the null and alternative hypotheses.

Exercise:

Problem:

The American family has an average of two children. What is the random variable? Describe in words.

Solution:

The random variable is the mean number of children an American family has.

Exercise:

Problem:

The mean entry level salary of an employee at a company is $58,000. You believe it is higher for IT professionals in the company. State the null and alternative hypotheses.

Exercise:

Problem:

A sociologist claims the probability that a person picked at random in Times Square in New York City is visiting the area is 0.83. You want to test to see if the proportion is actually less. What is the random variable? Describe in words.

Solution:

The random variable is the proportion of people picked at random in Times Square visiting the city.

Exercise:

Problem:

A sociologist claims the probability that a person picked at random in Times Square in New York City is visiting the area is 0.83. You want to test to see if the claim is correct. State the null and alternative hypotheses.

Exercise:

Problem:

In a population of fish, approximately 42% are female. A test is conducted to see if, in fact, the proportion is less. State the null and alternative hypotheses.

Solution:

a. $H_0 : p = 0.42$
b. $H_a : p < 0.42$

Exercise:

Problem:

Suppose that a recent article stated that the mean time spent in jail by a first–time convicted burglar is 2.5 years. A study was then done to see if the mean time has increased in the new century. A random sample of 26 first-time convicted burglars in a recent year was picked. The mean length of time in jail from the survey was 3 years with a standard deviation of 1.8 years. Suppose that it is somehow known that the population standard deviation is 1.5. If you were conducting a hypothesis test to determine if the mean length of jail time has increased, what would the null and alternative hypotheses be? The distribution of the population is normal.

a. H_0 : ________
b. H_a : ________

Exercise:

Problem:

A random survey of 75 death row inmates revealed that the mean length of time on death row is 17.4 years with a standard deviation of 6.3 years. If you were conducting a hypothesis test to determine if the population mean time on death row could likely be 15 years, what would the null and alternative hypotheses be?

a. H_0 : ___________
b. H_a : ___________

Solution:

a. $H_0 : \mu = 15$
b. $H_a : \mu \neq 15$

Exercise:

Problem:

The National Institute of Mental Health published an article stating that in any one-year period, approximately 9.5 percent of American adults suffer from depression or a depressive illness. Suppose that in a survey of 100 people in a certain town, seven of them suffered from depression or a depressive illness. If you were conducting a hypothesis test to determine if the true proportion of people in that town suffering from depression or a depressive illness is lower than the percent in the general adult American population, what would the null and alternative hypotheses be?

a. H_0 : _________
b. H_a : _________

Homework

Exercise:

Problem:

Some of the following statements refer to the null hypothesis, some to the alternate hypothesis.

State the null hypothesis, H_0 , and the alternative hypothesis. H_a , in terms of the appropriate parameter (μ or p).

a. The mean number of years Americans work before retiring is 34.
b. More than 60% of Americans vote in presidential elections.
c. The mean starting salary for San Jose State University graduates is less than $100,000 per year.
d. Twenty-nine percent of high school seniors get drunk each month.
e. Fewer than 5% of adults ride the bus to work in Los Angeles.
f. The mean number of cars a person owns in her lifetime is more than ten.
g. About half of Americans prefer to live away from cities, given the choice.
h. Europeans have a mean paid vacation each year of six weeks.
i. The chance of developing breast cancer is under 11% for women.
j. Private universities' mean tuition cost is more than $20,000 per year.

Solution:

a. $H_0 : \mu = 34$; $H_a : \mu \neq 34$
b. $H_0 : p = 0.60$; $H_a : p > 0.60$
c. $H_0 : \mu = 100{,}000$; $H_a : \mu < 100{,}000$
d. $H_0 : p = 0.29$; $H_a : p \neq 0.29$
e. $H_0 : p = 0.05$; $H_a : p < 0.05$
f. $H_0 : \mu = 10$; $H_a : \mu > 10$
g. $H_0 : p = 0.50$; $H_a : p \neq 0.50$
h. $H_0 : \mu = 6$; $H_a : \mu \neq 6$
i. $H_0 : p = 0.11$; $H_a : p < 0.11$
j. $H_0 : \mu = 20{,}000$; $H_a : \mu > 20{,}000$

Exercise:

Problem:

Over the past few decades, public health officials have examined the link between weight concerns and teen girls' smoking. Researchers surveyed a group of 273 randomly selected teen girls living in Massachusetts (between 12 and 15 years old). After four years the girls were surveyed again. Sixty-three said they smoked to stay thin. Is there good evidence that more than thirty percent of the teen girls smoke to stay thin? The alternative hypothesis is:

a. $p < 0.30$
b. $p \neq 0.30$
c. $p \geq 0.30$
d. $p > 0.30$

Exercise:

Problem:

A statistics instructor believes that fewer than 20% of Evergreen Valley College (EVC) students attended the opening night midnight showing of the latest Harry Potter movie. She surveys 84 of her students and finds that 11 attended the midnight showing. An appropriate alternative hypothesis is:

a. $p = 0.20$
b. $p > 0.20$
c. $p < 0.20$
d. $p \leq 0.20$

Solution:

c

Exercise:

Problem:

Previously, an organization reported that teenagers spent 4.5 hours per week, on average, on the phone. The organization thinks that, currently, the mean is higher. Fifteen randomly chosen teenagers were asked how many hours per week they spend on the phone. The sample mean was 4.75 hours with a sample standard deviation of 2.0. Conduct a hypothesis test. The null and alternative hypotheses are:

a. $H_o : x = 4.5$, $H_a : x > 4.5$
b. $H_o : \mu \geq 4.5$, $H_a : \mu < 4.5$
c. $H_o : \mu = 4.75$, $H_a : \mu > 4.75$
d. $H_o : \mu = 4.5$, $H_a : \mu > 4.5$

References

Data from the National Institute of Mental Health. Available online at http://www.nimh.nih.gov/publicat/depression.cfm.

Glossary

Hypothesis
a statement about the value of a population parameter, in case of two hypotheses, the statement assumed to be true is called the null hypothesis (notation H_0) and the statement one is trying to find evidence to support is called the alternative hypothesis (notation H_a).

The null hypothesis
a statement about the population that either is believed to be true or is used to put forth an argument unless it can be shown to be incorrect beyond a reasonable doubt.

The alternative hypothesis
a claim about the population that we are looking for evidence to support and what we conclude when we reject the null the hypothesis.

Outcomes and the Type I and Type II Errors

When you perform a hypothesis test, there are four possible outcomes depending on the actual truth (or falseness) of the null hypothesis H_0 and the decision to reject or not. The outcomes are summarized in the following table:

ACTION	**H_0 IS ACTUALLY**	**...**
	True	False
Do not reject H_0	Correct Outcome	Type II error
Reject H_0	Type I Error	Correct Outcome

The four possible outcomes in the table are:

1. The decision is **not to reject H_0** when **H_0 is true** (a correct decision).
2. The decision is to **reject H_0** when **H_0 is true** (an incorrect decision known as a **Type I error**).
3. The decision is **not to reject H_0** when, in fact, **H_0 is false** (an incorrect decision known as a **Type II error**).
4. The decision is to **reject H_0** when **H_0 is false** (a correct decision whose probability is called the **Power of the Test**).

Note: When the null hypothesis is rejected, the conclusion is that there is evidence that the alternative hypothesis is true.

Each of the errors occurs with a particular probability. The Greek letters α and β represent the probabilities.

α = probability of a Type I error = $P(\text{Type I error})$ = probability of rejecting the null hypothesis when the null hypothesis is true.

β = probability of a Type II error = $P(\text{Type II error})$ = probability of not rejecting the null hypothesis when the null hypothesis is false.

α and β should be as small as possible because they are probabilities of errors. They are rarely zero.

The Power of the Test is 1 – β. Ideally, we want a high power that is as close to one as possible. Increasing the sample size can increase the Power of the Test.

The following are examples of Type I and Type II errors.

Example:
Suppose the null hypothesis, H_0 , is: Frank's rock climbing equipment is safe.
Type I error: Frank thinks that his rock climbing equipment may not be safe when, in fact, it really is safe. **Type II error**: Frank thinks that his rock climbing equipment may be safe when, in fact, it is not safe.
α = the probability that Frank thinks his rock climbing equipment may not be safe when, in fact, it really is safe. β = the probability that Frank thinks his rock climbing equipment may be safe when, in fact, it is not safe.
Notice that, in this case, the error with the greater consequence is the Type II error. (If Frank thinks his rock climbing equipment is safe, he will go ahead and use it.)

Note:
Try It
Exercise:

Problem:

Suppose the null hypothesis, H_0 , is: the blood cultures contain no traces of pathogen X. State the Type I and Type II errors.

Solution:

Type I error: The researcher thinks the blood cultures do contain traces of pathogen X, when in fact, they do not.

Type II error: The researcher thinks the blood cultures do not contain traces of pathogen X, when in fact, they do.

Example:

Suppose the null hypothesis, H_0 , is: The victim of an automobile accident is alive when he arrives at the emergency room of a hospital.

Type I error: The emergency crew thinks that the victim is dead when, in fact, the victim is alive. **Type II error**: The emergency crew does not know if the victim is alive when, in fact, the victim is dead.

α = the probability that the emergency crew thinks the victim is dead when, in fact, he is really alive = $P(\text{Type I error})$. β = the probability that the emergency crew does not know if the victim is alive when, in fact, the victim is dead = $P(\text{Type II error})$.

The error with the greater consequence is the Type I error. (If the emergency crew thinks the victim is dead, they will not treat him.)

Note:

Try It

Exercise:

Problem:

Suppose the null hypothesis, H_0 , is: a patient is not sick. Which type of error has the greater consequence, Type I or Type II?

Solution:

The error with the greater consequence is the Type II error: the patient will be thought well when, in fact, he is sick, so he will not get treatment.

Example:
It's a Boy Genetic Labs claim to be able to increase the likelihood that a pregnancy will result in a boy being born. Statisticians want to test the claim. Suppose that the null hypothesis, H_0 , is: It's a Boy Genetic Labs has no effect on gender outcome.

Type I error: This results when a true null hypothesis is rejected. In the context of this scenario, we would state that we believe that It's a Boy Genetic Labs influences the gender outcome, when in fact it has no effect. The probability of this error occurring is denoted by the Greek letter alpha, α.

Type II error: This results when we fail to reject a false null hypothesis. In context, we would state that It's a Boy Genetic Labs does not influence the gender outcome of a pregnancy when, in fact, it does. The probability of this error occurring is denoted by the Greek letter beta, β.

The error of greater consequence would be the Type I error since couples would use the It's a Boy Genetic Labs product in hopes of increasing the chances of having a boy.

Note:
Try It
Exercise:

Problem:

“Red tide” is a bloom of poison-producing algae–a few different species of a class of plankton called dinoflagellates. When the weather and water conditions cause these blooms, shellfish such as clams living in the area develop dangerous levels of a paralysis-inducing toxin. In Massachusetts, the Division of Marine Fisheries (DMF) monitors levels of the toxin in shellfish by regular sampling of shellfish along the coastline. If the mean level of toxin in clams exceeds 800 μg (micrograms) of toxin per kg of clam meat in any area, clam harvesting is banned there until the bloom is over and levels of toxin in clams subside. The hypotheses used for this test are

$H_0 : \mu = 800$ micrograms and $H_a : \mu > 800$ micrograms

Describe both a Type I and a Type II error in this context, and state which error has the greater consequence.

Solution:

Type I error: The DMF believes that toxin levels are still too high when, in fact, toxin levels are at 800 micrograms. The DMF continues the harvesting ban.

Type II error: The DMF believes that toxin levels are within acceptable levels (are at 800 micrograms) when, in fact, toxin levels are still too high (more than 800 micrograms). The DMF lifts the harvesting ban. This error could be the most serious. If the ban is lifted and clams are still toxic, consumers could possibly eat tainted food.

In summary, the more dangerous error would be to commit a Type II error, because this error involves the availability of tainted clams for consumption.

Example:

A certain experimental drug claims a cure rate of at least 75% for males with prostate cancer. A cancer patient is skeptical and tests this claim using hypotheses

$H_0 : p = 0.75$ and $H_a : p < 0.75$

Describe both the Type I and Type II errors in context. Which error is the more serious?

Type I: A cancer patient believes the cure rate for the drug is less than 75% when it actually is 75%.

Type II: A cancer patient believes the experimental drug has a 75% cure rate when it has a cure rate that is less than 75%.

In this scenario, the Type II error contains the more severe consequence. If a patient believes the drug works 75% of the time, this most likely will influence the patient’s (and doctor’s) choice about whether to use the drug as a treatment option.

Note:

Try It

Determine both Type I and Type II errors for the following scenario: Assume a null hypothesis, H_0 , that states the percentage of adults with jobs is 88%.

Exercise:

Problem:

Identify the Type I and Type II errors from these four statements.

a. Not to reject the null hypothesis that the percentage of adults who have jobs is 88% when that percentage is actually less than 88%
b. Not to reject the null hypothesis that the percentage of adults who have jobs is 88% when the percentage is actually 88%.
c. Reject the null hypothesis that the percentage of adults who have jobs is 88% when the percentage is actually 88%.

d. Reject the null hypothesis that the percentage of adults who have jobs is when that percentage is actually less than 88%.

Solution:

Type I error: c

Type II error: a

Chapter Review

In every hypothesis test, the outcomes are dependent on a correct interpretation of the data. Incorrect calculations or misunderstood summary statistics can yield errors that affect the results. A **Type I** error occurs when a true null hypothesis is rejected. A **Type II error** occurs when a false null hypothesis is not rejected.

The probabilities of these errors are denoted by the Greek letters α and β, for a Type I and a Type II error respectively. The power of the test, 1 – β, quantifies the likelihood that a test will yield the correct result of a true alternative hypothesis being accepted. A high power is desirable.

Formula Review

α = the probability of a Type I error = $P(\text{Type I error})$ = the probability of rejecting the null hypothesis when the null hypothesis is true.

β = the probability of a Type II error = $P(\text{Type II error})$ = the probability of not rejecting the null hypothesis when the null hypothesis is false.

Exercise:

Problem:

The mean price of mid-sized cars in a region is $32,000. A test is conducted to see if the claim is true. State the Type I and Type II errors in complete sentences.

Solution:

Type I: The mean price of mid-sized cars is $32,000, but we conclude that it is not $32,000.

Type II: The mean price of mid-sized cars is not $32,000, but we conclude that it is $32,000.

Exercise:

Problem:

A sleeping bag is tested to withstand temperatures of –15 °F. You think the bag cannot stand temperatures that low. State the Type I and Type II errors in complete sentences.

Exercise:

Problem: For [link], what are α and β in words?

Solution:

α = the probability that you think the bag cannot withstand -15 degrees F, when in fact it can

β = the probability that you think the bag can withstand -15 degrees F, when in fact it cannot

Exercise:

Problem: In words, describe $1 – \beta$ for [link].

Exercise:

Problem:

A group of doctors is deciding whether or not to perform an operation. Suppose the null hypothesis, H_0 , is: the surgical procedure will go well. State the Type I and Type II errors in complete sentences.

Solution:

Type I: The procedure will go well, but the doctors think it will not.

Type II: The procedure will not go well, but the doctors think it will.

Exercise:

Problem:

A group of doctors is deciding whether or not to perform an operation. Suppose the null hypothesis, H_0 , is: the surgical procedure will go well. Which is the error with the greater consequence?

Exercise:

Problem:

The power of a test is 0.981. What is the probability of a Type II error?

Solution:

0.019

Exercise:

Problem:

A group of divers is exploring an old sunken ship. Suppose the null hypothesis, H_0 , is: the sunken ship does not contain buried treasure. State the Type I and Type II errors in complete sentences.

Exercise:

Problem:

A microbiologist is testing a water sample for E-coli. Suppose the null hypothesis, H_0 , is: the sample does not contain E-coli. The probability that the sample does not contain E-coli, but the microbiologist thinks it does is 0.012. The probability that the sample does contain E-coli, but the microbiologist thinks it does not is 0.002. What is the power of this test?

Solution:

0.998

Exercise:

Problem:

A microbiologist is testing a water sample for E-coli. Suppose the null hypothesis, H_0 , is: the sample contains E-coli. Which is the error with the greater consequence?

Homework

Exercise:

Problem:

State the Type I and Type II errors in complete sentences given the following statements, which may represent a null or alternative hypothesis.

a. The mean number of years Americans work before retiring is 34.
b. More than 60% of Americans vote in presidential elections.
c. The mean starting salary for San Jose State University graduates is less than $100,000 per year.
d. Twenty-nine percent of high school seniors get drunk each month.
e. Fewer than 5% of adults ride the bus to work in Los Angeles.

f. The mean number of cars a person owns in his or her lifetime is less than ten.
g. About half of Americans prefer to live away from cities, given the choice.
h. Europeans have a mean paid vacation each year of six weeks.
i. The chance of developing breast cancer is under 11% for women.
j. Private universities mean tuition cost is more than $20,000 per year.

Solution:

a. Type I error: We conclude that the mean is not 34 years, when it really is 34 years. Type II error: We believe that the mean is 34 years, when in fact it really is not 34 years.
b. Type I error: We conclude that more than 60% of Americans vote in presidential elections, when the actual percentage is 60%. Type II error: We believe that 60% of Americans vote in presidential elections when, in fact, more than 60% do.
c. Type I error: We conclude that the mean starting salary is less than $100,000, when it really is $100,000. Type II error: We believe that the mean starting salary is $100,000 when, in fact, it is less than $100,000.
d. Type I error: We conclude that the proportion of high school seniors who get drunk each month is not 29%, when it really is 29%. Type II error: We believe that the proportion of high school seniors who get drunk each month is 29% when, in fact, it is not 29%.
e. Type I error: We conclude that fewer than 5% of adults ride the bus to work in Los Angeles, when the percentage that do is really 5%. Type II error: We believe that 5% of adults ride the bus to work in Los Angeles when, in fact, fewer than 5% do.
f. Type I error: We conclude that the mean number of cars a person owns in his or her lifetime is less than 10, when in reality it is 10. Type II error: We believe that the mean number of cars a person owns in his or her lifetime is 10 when, in fact, it is less than 10.

g. Type I error: We conclude that the proportion of Americans who prefer to live away from cities is not about half, though the actual proportion is about half. Type II error: We believe that the proportion of Americans who prefer to live away from cities is half when, in fact, it is not half.
h. Type I error: We conclude that the duration of paid vacations each year for Europeans is not six weeks, when in fact it is six weeks. Type II error: We believe that the duration of paid vacations each year for Europeans is six weeks when, in fact, it is not.
i. Type I error: We conclude that the proportion is less than 11%, when it is really 11%. Type II error: We believe that the proportion of women who develop breast cancer is 11%, when in fact it is less than 11%.
j. Type I error: We conclude that the average tuition cost at private universities is more than $20,000, though in reality it is $20,000. Type II error: We believe that the average tuition cost at private universities is $20,000 when, in fact, it is more than $20,000.

Exercise:

Problem:

For statements a-j in the previous exercise, answer the following in complete sentences.

a. State a consequence of committing a Type I error.
b. State a consequence of committing a Type II error.

Exercise:

Problem:

When a new drug is created, the pharmaceutical company must subject it to testing before receiving the necessary permission from the Food and Drug Administration (FDA) to market the drug. Suppose the null hypothesis is “the drug is unsafe.” What is the Type II Error?

a. To conclude the drug is safe when in, fact, it is unsafe.

b. Not to conclude the drug is safe when, in fact, it is safe.
c. To conclude the drug is safe when, in fact, it is safe.
d. Not to conclude the drug is unsafe when, in fact, it is unsafe.

Solution:

b

Exercise:

Problem:

A statistics instructor believes that fewer than 20% of Evergreen Valley College (EVC) students attended the opening midnight showing of the latest Harry Potter movie. She surveys 84 of her students and finds that 11 of them attended the midnight showing. The Type I error is to conclude that the percent of EVC students who attended is ________.

a. at least 20%, when in fact, it is less than 20%.
b. 20%, when in fact, it is 20%.
c. less than 20%, when in fact, it is 20%.
d. less than 20%, when in fact, it is less than 20%.

Exercise:

Problem:

It is believed that Lake Tahoe Community College (LTCC) Intermediate Algebra students get less than seven hours of sleep per night, on average. A survey of 22 LTCC Intermediate Algebra students generated a mean of 7.24 hours with a standard deviation of 1.93 hours. At a level of significance of 5%, do LTCC Intermediate Algebra students get less than seven hours of sleep per night, on average?

The Type II error is not to reject that the mean number of hours of sleep LTCC students get per night is seven hours when, in fact, the mean number of hours

a. is more than seven hours.
b. is at most seven hours.
c. is at least seven hours.
d. is less than seven hours.

Solution:

d

Exercise:

Problem:

Previously, an organization reported that teenagers spent 4.5 hours per week, on average, on the phone. The organization thinks that, currently, the mean is higher. Fifteen randomly chosen teenagers were asked how many hours per week they spend on the phone. The sample mean was 4.75 hours with a sample standard deviation of 2.0. Conduct a hypothesis test, the Type I error is:

a. to conclude that the current mean hours per week is higher than 4.5, when in fact, it is higher
b. to conclude that the current mean hours per week is higher than 4.5, when in fact, it is the same
c. to conclude that the mean hours per week currently is 4.5, when in fact, it is higher
d. to conclude that the mean hours per week currently is no higher than 4.5, when in fact, it is not higher

Glossary

Type 1 Error
: The decision is to reject the null hypothesis when, in fact, the null hypothesis is true.

Type 2 Error

The decision is not to reject the null hypothesis when, in fact, the null hypothesis is false.

Distribution Needed for Hypothesis Testing

Earlier in the course, we discussed sampling distributions. **Particular distributions are associated with hypothesis testing.** We perform tests of a population mean using a **normal distribution** or a **Student's t-distribution**. (Remember, use a Student's t-distribution when the population **standard deviation** is unknown and the sampling distribution of the sample mean is approximately normal.) We perform tests of a population proportion using a normal distribution (usually n, the sample size, is large).

If you are testing a **single population mean**, the distribution for the test is for **means**:

$X \sim N\left(\mu_x, \frac{\sigma_x}{\sqrt{n}}\right)$, when σ_x is known, or t_{df}, when σ_x is unknown

The population parameter is μ.

The estimated value (point estimate) for μ is x, the sample mean.

If you are testing a **single population proportion**, the distribution for the test is for proportions or percentages:

$$\widehat{P} \sim N\left(p, \sqrt{\frac{p \cdot q}{n}}\right)$$

The population parameter is p.

The estimated value (point estimate) for p is $\widehat{p}$, the sample proportion.

$\widehat{p} = \frac{x}{n}$ where x is the number of successes in the sample and n is the sample size.

Assumptions

When you perform a **hypothesis test of a single population mean** μ using a **Student's t-distribution** (often called a t-test), there are fundamental assumptions that need to be met in order for the test to be valid. Your data should be a **simple random sample** that comes from a population that is approximately **normally distributed**. You use the sample **standard deviation** to approximate the population standard deviation. (Note that if the sample size is sufficiently large (at least 30), a t-test will work even if the population is not approximately normally distributed).

When you perform a **hypothesis test of a single population mean** μ using a normal distribution (often called a z-test), you take a simple random sample from the population. The population you are testing is normally distributed or your sample size is sufficiently large (at least 30). You know the value of the population standard deviation which, in reality, is rarely known.

When you perform a **hypothesis test of a single population proportion** p, you take a simple random sample from the population. You must meet the conditions for a **binomial distribution** which are: there are a certain number n of independent trials, the outcomes of any trial are success or failure, and each trial has the same probability of a success p. The shape of the binomial distribution needs to be similar to the shape of the normal distribution. To ensure this, the quantities np (the expected number of successes, assuming the null hypothesis is true) and nq (the expected number of failures, assuming the null hypothesis is true) must both be at least 10 ($np \geq 10$ and $nq \geq 10$). Then the binomial distribution of a sample (estimated) proportion can be approximated by the normal distribution with $\mu = p$ and $\sigma = \sqrt{\frac{pq}{n}}$. Remember that $q = 1 - p$.

Section Review

In order for a hypothesis test’s results to be generalized to a population, certain requirements must be satisfied.

When testing for a single population mean:

1. A Student's t-test should be used if the data come from a simple, random sample and the population is approximately normally distributed, or the sample size is large (at least 30), with an unknown standard deviation.
2. The normal test will work if the data come from a simple, random sample and the population is approximately normally distributed, or the sample size is large (at least 30), with a known standard deviation.

When testing a single population proportion use a normal test for a single population proportion if the data comes from a simple, random sample, fill the requirements for a binomial distribution, and the mean number of success and the mean number of failures satisfy the conditions: $np \geq 10$ and $nq \geq 10$ where n is the sample size, p is the probability of a success, and q is the probability of a failure.

Formula Review

Types of Hypothesis Tests

- Single population mean, **known** population standard deviation: **Normal test**.
- Single population mean, **unknown** population standard deviation: **Student's t-test**.
- Single population proportion: **Normal test**.
- For a **single population mean**, we may use a normal distribution with the following mean and standard deviation. Means: $\mu = \mu_x$ and $\sigma = \frac{\sigma_x}{\sqrt{n}}$
- A **single population proportion**, we may use a normal distribution with the following mean and standard deviation. Proportions: $\mu = p$ and $\sigma = \sqrt{\frac{pq}{n}}$.

Exercise:

Problem:

Which two distributions can you use for hypothesis testing for this chapter?

Solution:

A normal distribution or a Student’s t-distribution

Exercise:

Problem:

Which distribution do you use when you are testing a population mean and the standard deviation is known? Assume the sample size is large.

Exercise:

Problem:

Which distribution do you use when the standard deviation is not known and you are testing one population mean? Assume the sample size is large.

Solution:

Use a Student’s t-distribution

Exercise:

Problem:

A sample mean is 12.8, and the sample standard deviation is two. The sample size is 20. What distribution should you use to perform a hypothesis test? Assume the underlying population is normal.

Exercise:

Problem:

A population has a standard deviation of five. The sample mean is 24, and the sample size is 108. What distribution should you use to perform a hypothesis test?

Solution:

A normal distribution for a single population mean

Exercise:

Problem:

It is thought that 42% of respondents in a taste test would prefer Brand *A*. In a particular test of 100 people, 39% preferred Brand *A*. What distribution should you use to perform a hypothesis test?

Exercise:

Problem:

You are performing a hypothesis test of a single population mean using a Student's t-distribution. What must you assume about the distribution of the data?

Solution:

It must be approximately normally distributed.

Exercise:

Problem:

You are performing a hypothesis test of a single population mean using a Student’s t-distribution. The data are not from a simple random sample. Can you accurately perform the hypothesis test?

Exercise:

Problem:

You are performing a hypothesis test of a single population proportion. What must be true about the quantities of np and nq?

Solution:

They must both be at least ten.

Exercise:

Problem:

You are performing a hypothesis test of a single population proportion. You find out that np is less than ten. What must you do to be able to perform a valid hypothesis test?

Exercise:

Problem:

You are performing a hypothesis test of a single population proportion. The data come from which distribution?

Solution:

binomial distribution

Homework

Exercise:

Problem:

It is believed that Lake Tahoe Community College (LTCC) Intermediate Algebra students get less than seven hours of sleep per night, on average. A survey of 22 LTCC Intermediate Algebra students generated a mean of 7.24 hours with a standard deviation of 1.93 hours. At a level of significance of 5%, do LTCC Intermediate Algebra students get less than seven hours of sleep per night, on average? The distribution to be used for this test is X ~ ________________

a. $N(7.24, \frac{1.93}{\sqrt{22}})$
b. $N(7.24, 1.93)$
c. t_{22}
d. t_{21}

Solution:

d

Rare Events, the Sample, Decision, and Conclusion

Establishing the type of distribution, sample size, and known or unknown standard deviation can help you figure out how to go about a hypothesis test. However, there are several other factors you should consider when working out a hypothesis test.

Rare Events

Suppose you make an assumption about a property of the population (this assumption is the **null hypothesis**). Then you gather sample data randomly. If the sample has properties that would be very **unlikely** to occur if the assumption is true, then you would conclude that your assumption about the population is probably incorrect. (Remember that your assumption is just an **assumption**—it is not a fact and it may or may not be true. But your sample data are real and the data are showing you a fact that seems to contradict your assumption.)

For example, Didi and Ali are at a birthday party of a very wealthy friend. They hurry to be first in line to grab a prize from a tall basket that they cannot see inside of because they will be blindfolded. There are 200 plastic bubbles in the basket and Didi and Ali have been told that there is only one with a $100 bill. Didi is the first person to reach into the basket and pull out a bubble. Her bubble contains a $100 bill. The probability of this happening is $\frac{1}{200} = 0.005$. Because this is so unlikely, Ali is hoping that what the two of them were told is wrong and there are more $100 bills in the basket. A "rare event" has occurred (Didi getting the $100 bill) so Ali doubts the assumption about only one $100 bill being in the basket.

Using the Sample to Test the Null Hypothesis

Use the sample data to calculate the actual probability of getting the test result, called the ***p*-value**. The *p*-value is the **probability that, if the null hypothesis is true, the results from another randomly selected sample**

will be as extreme or more extreme as the results obtained from the given sample.

A large p-value calculated from the data indicates that we should not reject the **null hypothesis**. The smaller the p-value, the more unlikely the outcome, and the stronger the evidence is against the null hypothesis. We would reject the null hypothesis if the evidence is strongly against it.

Tip: Draw a graph that shows the p-value. The hypothesis test is easier to perform if you use a graph because you see the problem more clearly.

Example:

Suppose a baker claims that his bread height is more than 15 cm, on average. Several of his customers do not believe him. To persuade his customers that he is right, the baker decides to do a hypothesis test. He bakes 10 loaves of bread. The mean height of the sample loaves is 17 cm. The baker knows from baking hundreds of loaves of bread that the **standard deviation** for the height is 0.5 cm. and the distribution of heights is normal.

The null hypothesis is $H_0 : \mu = 15$ and the alternate hypothesis is $H_a : \mu > 15$, since the baker is trying to show evidence that the heights of his loaves are **more than** 15 cm on **average**.

The words **"is more than"** translates as a ">" so "$\mu > 15$" is the alternate hypothesis. The null hypothesis must always contain the equal sign.

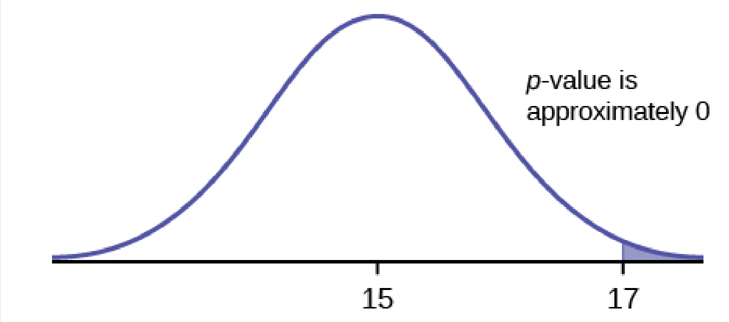
Since σ **is known** ($\sigma = 0.5$ cm.) and the distribution for the population is known to be normal, the distribution for this test is also normal with standard deviation $\frac{\sigma}{\sqrt{n}} = \frac{0.5}{\sqrt{10}} = 0.16$.

Suppose the null hypothesis is true (the mean height of the loaves is 15 cm, $\mu = 15$). Then, is the mean height (17 cm) calculated from the sample unexpectedly large?

The hypothesis test works by asking the question how **unlikely** the sample mean would be **if the null hypothesis were true**.

The graph shows how far out the sample mean is on the normal curve. The p-value is the probability that, if we were to take other samples, any other sample mean would fall at least as far out as 17 cm.

The p-value is the probability that a sample mean is the same or greater than 17 cm when the population mean is, in fact, 15 cm. We can calculate this probability using the normal distribution for means.



The p-value = $P(x > 17)$ which is approximately zero. (Use normalcdf on your calculator.)

A p-value of approximately zero tells us that it is highly unlikely that a loaf of bread rises only 15 cm, on average. That is, almost 0% of all loaves of bread would be at least as high as 17 cm **purely by CHANCE** had the population mean height really been 15 cm. Because the outcome of 17 cm is so **unlikely (meaning it is happening NOT by chance alone)**, we conclude that the evidence is strongly against the null hypothesis (that the mean height is 15 cm). There is sufficient evidence that the true mean height for the population of the baker's loaves of bread is greater than 15 cm.

Note:

Try It

Exercise:

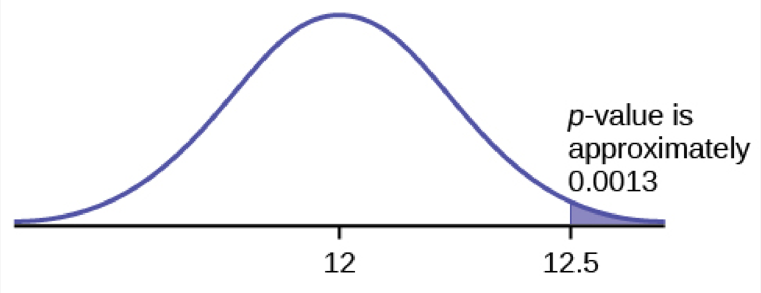
Problem:

A population has a normal distribution with an unknown mean and a standard deviation of 1. We want to verify a claim that the mean is greater than 12. A sample of 36 is taken with a sample mean of 12.5.

$H_0 : \mu = 12$
$H_a : \mu > 12$
The *p*-value is 0.0013
Draw a graph that shows the *p*-value.

Solution:

p-value = 0.0013



Decision and Conclusion

A systematic way to make a decision of whether to reject or not reject the **null hypothesis** is to compare the *p*-value to a **preset or preconceived** α **(the greek letter alpha, also called a "significance level")**. The significance level may or may not be given to you at the beginning of the problem.

When you make a **decision** to reject or not reject H_0 , do as follows:

- If the p-value $\leq \alpha$, reject H_0 . The results of the sample data are significant. There is sufficient evidence to conclude that the **alternative hypothesis**, H_a , is true.
- If the p-value $> \alpha$, do not reject H_0 . The results of the sample data are not significant. There is not sufficient evidence to conclude that the alternative hypothesis,H_a , is true.
- Note: When you "do not reject H_0", it does not mean that you should conclude that H_0 is true. It simply means that the sample data have **failed** to provide sufficient evidence to cast serious doubt about the truthfulness of H_0 .

Conclusion: After you make your decision, write a thoughtful conclusion about the hypotheses **in terms of the given problem**.

Example:

When using the p-value to evaluate a hypothesis test, it is sometimes useful to use the following memory device

If the p-value is low, the null must go.

If the p-value is high, the null must fly.

This memory aid relates a p-value less than the established alpha (the p is low) as rejecting the null hypothesis and, likewise, relates a p-value higher than the established alpha (the p is high) as not rejecting the null hypothesis.

Exercise:

Problem: Fill in the blanks.

Reject the null hypothesis when ____.

The results of the sample data ____.

Do not reject the null when hypothesis when ____.

The results of the sample data ____.

Solution:

Reject the null hypothesis when **the** ***p*****-value is less than the established alpha value**. The results of the sample data **support the alternative hypothesis**.

Do not reject the null hypothesis when **the** ***p*****-value is greater than the established alpha value**. The results of the sample data **do not support the alternative hypothesis**.

Note:
Try It
Exercise:

Problem:

"It’s a Boy Genetics Labs" claim their procedures improve the chances of a boy being born. The results for a test of a single population proportion are as follows:

$H_0 : p = 0.50$, $H_a : p > 0.50$

$\alpha = 0.01$

p-value = 0.025

Interpret the results and state a conclusion in simple, non-technical terms.

Solution:

Since the p-value is greater than the established alpha value (the p-value is high), we do not reject the null hypothesis. There is not enough evidence to support "It’s a Boy Genetics Labs" stated claim that their procedures improve the chances of a boy being born.

Chapter Review

When the probability of an event occurring is low, and it happens, it is called a rare event. Rare events are important to consider in hypothesis testing because they can inform your willingness not to reject or to reject a null hypothesis. To test a null hypothesis, find the p-value for the sample data and graph the results. When deciding whether or not to reject the null the hypothesis, keep these two parameters in mind:

1. If the p-value $\le \alpha$, reject the null hypothesis
2. If p-value $> \alpha$, do not reject the null hypothesis

Exercise:

Problem: When do you reject the null hypothesis?

Exercise:

Problem:

The probability of winning the grand prize at a particular carnival game is 0.005. Is the outcome of winning very likely or very unlikely?

Solution:

The outcome of winning is very unlikely.

Exercise:

Problem:

The probability of winning the grand prize at a particular carnival game is 0.005. Michele wins the grand prize. Is this considered a rare or common event? Why?

Exercise:

Problem:

It is believed that the mean height of high school students who play basketball on the school team is 73 inches with a standard deviation of 1.8 inches. A claim is made that the mean height is actually less than 73 inches. A random sample of 40 players is chosen. The sample mean was 71 inches, and the sample standard deviation was 1.5 years. Do the data support the claim that the mean height is less than 73 inches? The p-value is almost zero. State the null and alternative hypotheses and interpret the p-value.

Solution:

$H_0 : \mu = 73$
$H_a : \mu < 73$
The p-value is almost zero, which means there is sufficient evidence, at the 5% level, to conclude that the mean height of high school students who play basketball on the school team is less than 73 inches. The data do support the claim.

Exercise:

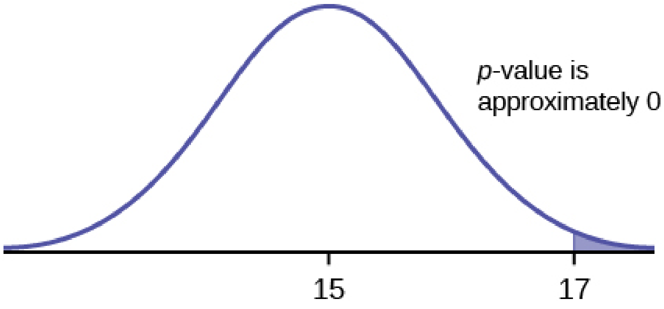
Problem:

The mean age of graduate students at a University is claimed to be more than 31 years with a known standard deviation of two years. A random sample of 15 graduate students is taken. (Assume the population is normal.) The sample mean is 32 years and the sample standard deviation is three years. Are the data significant at the 1% level? The p-value is 0.0264. State the null and alternative hypotheses and interpret the p-value.

Exercise:

Problem:

Does the shaded region represent a low or a high p-value compared to a level of significance of 1%?



Solution:

The shaded region shows a low p-value.

Exercise:

Problem: What should you do when the p-value < α?

Exercise:

Problem: What should you do if the p-value > α?

Solution:

Do not reject H_0.

Exercise:

Problem:

If you do not reject the null hypothesis, then it must be true. Is this statement correct? State why or why not in complete sentences.

Use the following information to answer the next seven exercises: Suppose that a recent article stated that the mean time spent in jail by a first-time convicted burglar is 2.5 years. A study was then done to see if the mean time has increased in the new century. A random sample of 26 first-time convicted burglars in a recent year was picked. The mean length of time in jail from the survey was three years with a standard deviation of 1.8 years. Suppose that it is somehow known that the population standard deviation is 1.5 years. Conduct a hypothesis test to determine if the mean length of jail time has increased. Assume the distribution of the jail times is approximately normal.

Exercise:

Problem: Is this a test of means or proportions?

Solution:

means

Exercise:

Problem: What symbol represents the random variable for this test?

Exercise:

Problem: In words, define the random variable for this test.

Solution:

the mean time spent in jail for 26 first time convicted burglars

Exercise:

Problem:

Is the population standard deviation known and, if so, what is it?

Exercise:

Problem: Calculate the following:

a. x ________
b. σ ________
c. s_x ________
d. n________

Solution:

a. 3
b. 1.5
c. 1.8
d. 26

Exercise:

Problem:

Since both σ and s_x are given, which should be used? In one to two complete sentences, explain why.

Exercise:

Problem: State the distribution to use for the hypothesis test.

Solution:

$$X \sim N\left(2.5, \frac{1.5}{\sqrt{26}}\right)$$

Exercise:

Problem:

A random survey of 75 death row inmates revealed that the mean length of time on death row is 17.4 years with a standard deviation of 6.3 years. Conduct a hypothesis test to determine if the population mean time on death row could likely be 15 years.

a. Is this a test of one mean or proportion?
b. State the null and alternative hypotheses.
   H_0 : ______________________ H_a : ______________________
c. What symbol represents the random variable for this test?
d. In words, define the random variable for this test.
e. Is the population standard deviation known and, if so, what is it?
f. Calculate the following:

   i. x = ______________
   ii. s = ______________
   iii. n = ______________

g. Which test should be used?
h. State the distribution to use for the hypothesis test.
i. Find the p-value.
j. At a pre-conceived α = 0.05, what is your:

   i. Decision:
   ii. Reason for the decision:
   iii. Conclusion (write out in a complete sentence):

Homework

Exercise:

Problem:

The National Institute of Mental Health published an article stating that in any one-year period, approximately 9.5 percent of American adults suffer from depression or a depressive illness. Suppose that in a survey of 100 people in a certain town, seven of them suffered from depression or a depressive illness. Conduct a hypothesis test to determine if the true proportion of people in that town suffering from depression or a depressive illness is lower than the percent in the general adult American population.

a. Is this a test of one mean or proportion?
b. State the null and alternative hypotheses.
   H_0 : ______________________ H_a : ______________________
c. What symbol represents the random variable for this test?
d. In words, define the random variable for this test.
e. Calculate the following:

   i. x = __________________
   ii. n = __________________
   iii. $\widehat{p}$ = ______________

f. Calculate σ = ___________. Show the formula set-up.
g. State the distribution to use for the hypothesis test.
h. Find the p-value.
i. At a pre-conceived $\alpha = 0.05$, what is your:

   i. Decision:
   ii. Reason for the decision:
   iii. Conclusion (write out in a complete sentence):

Solution:

a. proportion
b. $H_0 : p = 0.095$ $H_a : p < 0.095$
c. $\widehat{P}$

d. $\widehat{P}$ = the proportion of people who've suffered from a depressive illness, in a sample of 100 people from a certain town.

e.
   i. x = 7
   ii. n = 100
   iii. $\widehat{p}$ = 0.07

f. $\sigma = \sqrt{\frac{(0.07)(1-0.07)}{100}} \approx 0.0255$
g. normal distribution for proportions
h. p-value = 0.1969

i.
   i. Do not reject the null hypothesis.
   ii. The p-value > 0.05.
   iii. There is insufficient evidence to conclude that the proportion of people in this town who've suffered a depressive illness is lower than the national proportion.

Glossary

Level of Significance of the Test

In hypothesis testing, the Level of Significance is called the preconceived α or the preset α.

p-value

the probability that an event (as in the result of a sample) will happen purely by chance assuming the null hypothesis is true. The smaller the p-value, the stronger the evidence is against the null hypothesis.

Additional Information and Full Hypothesis Test Examples

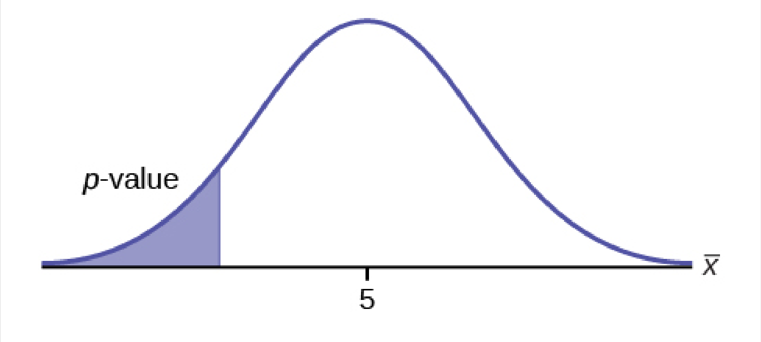
- In a **hypothesis test** problem, you may see words such as "the level of significance is 1%." The "1%" is the preconceived or preset α.
- The statistician setting up the hypothesis test selects the value of α to use **before** collecting the sample data.
- **If no level of significance is given, a common standard to use is α = 0.05.**
- When you calculate the p-value and draw the picture, the p-value is the area in the left tail, the right tail, or split evenly between the two tails. For this reason, we call the hypothesis test left-, right-, or two-tailed.
- The **alternative hypothesis**, H_a , tells you if the test is left-, right-, or two-tailed. It is the **key** to conducting the appropriate test.
- H_a **never** has a symbol that contains an equal sign.
- **Thinking about the meaning of the p-value**: A data analyst (and anyone else) should have more confidence that he made the correct decision to reject the null hypothesis with a smaller p-value (for example, 0.001 as opposed to 0.04) even if using the 0.05 level for alpha. Similarly, for a large p-value such as 0.4, as opposed to a p-value of 0.056 (alpha = 0.05 is less than either number), a data analyst should have more confidence that she made the correct decision in not rejecting the null hypothesis. This makes the data analyst use judgment rather than mindlessly applying rules.
- A conclusion of finding evidence to support the alternative hypothesis is equivalent to stating that the findings are **statistically significant** or that a **statistically significant difference** has been found between the claimed value of the population parameter (from the null hypothesis) and its true value.

The following examples illustrate a left-, right-, and two-tailed test.

Example:

$H_o : \mu = 5$, $H_a : \mu < 5$

Test of a single population mean. H_a tells you the test is **left-tailed**. The picture of the p-value is as follows:



Note that the *p*-value is calculated based on the assumption that the null hypothesis is true. This is why the distribution is centered at $\mu = 5$.

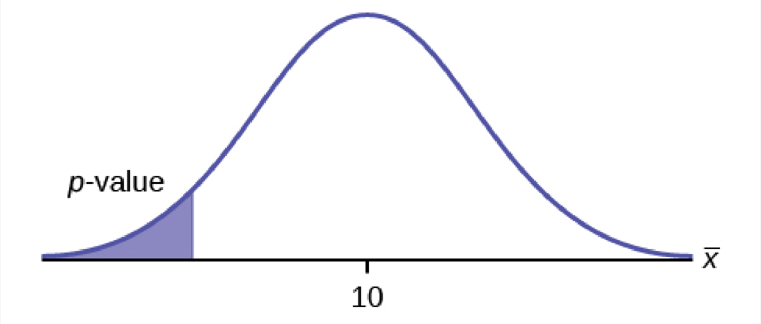
Note:
Try It
Exercise:

Problem: $H_0 : \mu = 10$, $H_a : \mu < 10$

Assume the *p*-value is 0.0935. What type of test is this? Draw the picture of the *p*-value.

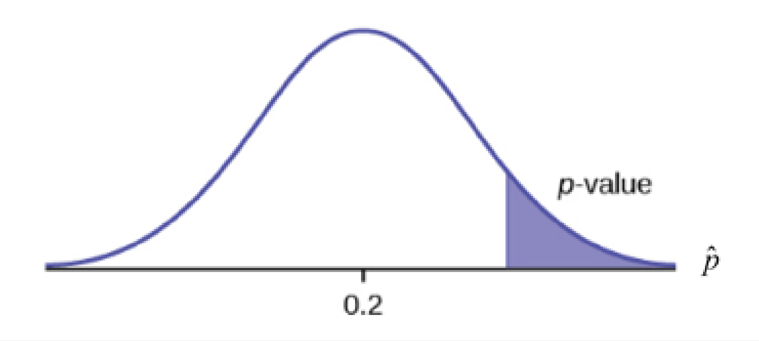
Solution:

left-tailed test



Example:
$H_0 : p = 0.2$ $H_a : p > 0.2$

This is a test of a single population proportion. H_a tells you the test is **right-tailed**. The picture of the *p*-value is as follows:



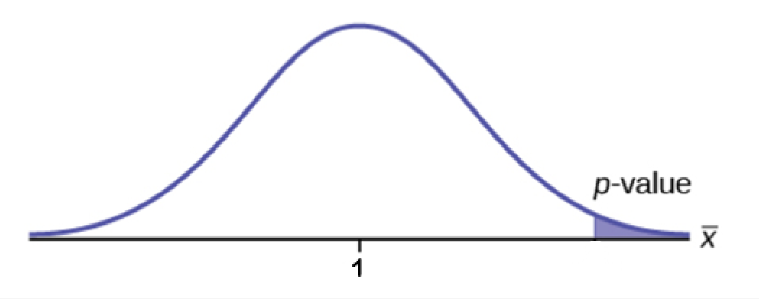
Note:
Try It
Exercise:

Problem: $H_0 : \mu = 1$, $H_a : \mu > 1$

Assume the *p*-value is 0.1243. What type of test is this? Draw the picture of the *p*-value.

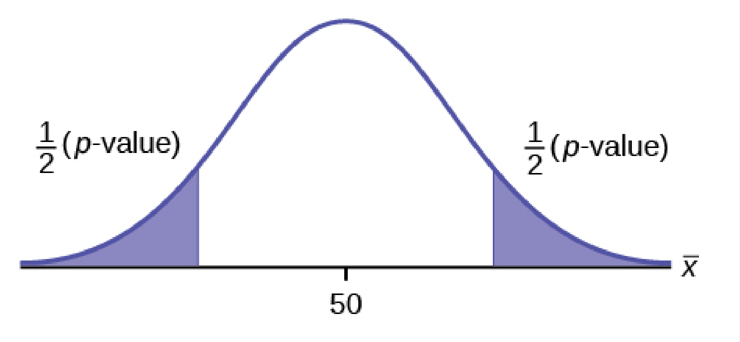
Solution:

right-tailed test



Example:
$H_0 : \mu = 50$ $H_a : \mu \neq 50$

This is a test of a single population mean. H_a tells you the test is **two-tailed**. The picture of the p-value is as follows.



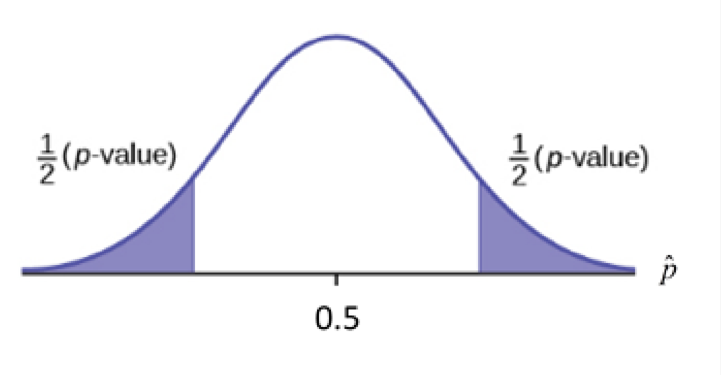
Note:
Try It
Exercise:

Problem: $H_0 : p = 0.5$, $H_a : p \neq 0.5$

Assume the p-value is 0.2564. What type of test is this? Draw the picture of the p-value.

Solution:

two-tailed test



Full Hypothesis Test Examples

Example:
Exercise:

Problem:

Jeffrey, as an eight-year old, established a **mean time of 16.43 seconds** for swimming the 25-yard freestyle, with a **standard deviation of 0.8 seconds**. His dad, Frank, thought that Jeffrey could swim the 25-yard freestyle faster using goggles. Frank bought Jeffrey a new pair of expensive goggles and timed Jeffrey for **fifteen 25-yard freestyle swims**. For the fifteen swims, **Jeffrey's mean time was 16 seconds. Frank thought that the goggles helped Jeffrey to swim faster than the 16.43 seconds.** Conduct a hypothesis test using a preset $\alpha = 0.05$. Assume that the swim times for the 25-yard freestyle are normal.

Solution:

Set up the Hypothesis Test:

Since the problem is about a mean, this is a **test of a single population mean**.

$H_0 : \mu = 16.43$ $\qquad$ $H_a : \mu < 16.43$

For Jeffrey to swim faster, his time will be less than 16.43 seconds. The "<" tells you this is left-tailed.

Determine the distribution needed:

Random variable: $\bar{X}$ = the mean time to swim the 25-yard freestyle, in a sample of 15 swims.

Distribution for the test: X is normal (population **standard deviation** is known: σ = 0.8)

$X \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ Therefore, $X \sim N\left(16.43, \frac{0.8}{\sqrt{15}}\right)$

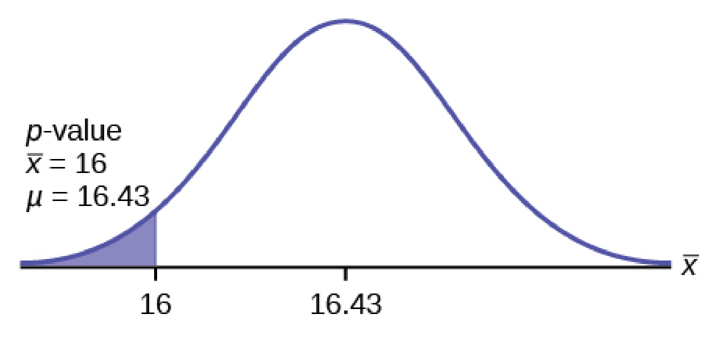
Remember, μ = 16.43 comes from H_0 and not the data; σ = 0.8, and n = 15.

Calculate the p-value using the normal distribution for a mean:

The p-value = $P(x < 16)$ = 0.0187, where the sample mean in the problem is given as 16.

The p-value = 0.0187. The p-value is the area to the left of the sample mean, which is given as 16.

Graph:



Interpretation of the p-value: If H_0 is true, there is a 0.0187 probability (1.87%) that Jeffrey's mean time to swim fifteen 25-yard freestyle swims would be 16 seconds or less. Because a 1.87% chance is small, the mean time of 16 seconds or less is unlikely to have happened randomly. It is a rare event.

Compare α and the p-value:

$\alpha = 0.05$ and p-value = 0.0187; p-value < α

Make a decision: Since p-value < α, reject H_0 .

This means that you reject that $\mu = 16.43$. In other words, you do not think Jeffrey swims the 25-yard freestyle in 16.43 seconds, on average, but instead he swims faster with the new goggles.

Conclusion: At the 5% significance level, we conclude that Jeffrey swims faster using the new goggles. The sample data show there is sufficient evidence that Jeffrey's mean time to swim the 25-yard freestyle, using the new goggles, is less than 16.43 seconds.

The p-value can easily be calculated using the z-Test on the calculator.

Note:
Press `STAT` and arrow over to `TESTS`. Press `1:Z-Test`. Arrow over to `Stats` and press `ENTER`. Arrow down and enter 16.43 for μ_0 (from the null hypothesis), .8 for σ, 16 for the sample mean, and 15 for n. Arrow down to μ : (alternate hypothesis) and arrow over to < μ_0, since $\mu < 16.43$ is the alternative hypothesis. Press `ENTER`. Arrow down to `Calculate` and press `ENTER`. The calculator not only calculates the p-value ($p = 0.0187$), but it also calculates the test statistic (z-score) for the sample mean. Do this set of instructions

again except arrow to `Draw`(instead of `Calculate`). Press `ENTER`. A shaded graph appears with z = -2.08 (test statistic) and p = 0.0187 (p-value). Make sure when you use `Draw` that no other equations are highlighted in Y = and the plots are turned off.

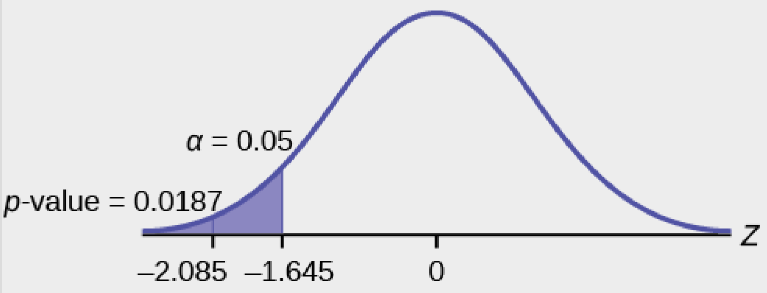
When the calculator does a Z-Test, the `Z-Test` function finds the p-value by doing a normal probability calculation using the **central limit theorem**:

$P(x<16) =$ `2nd DISTR normcdf` $\left(-1\text{E}99, 16, 16.43, 0.8/\sqrt{15}\right)$.

Note:

Historical Note

The traditional way to compare the two probabilities, α and the p-value, is to compare the critical value (z-score from α) to the test statistic (z-score from the sample data). The calculated test statistic for the p-value is –2.08. (From the Central Limit Theorem, the test statistic formula is $z = \frac{x - \mu_x}{\left(\frac{\sigma_x}{\sqrt{n}}\right)}$. For the previous example, x = 16, μ_x= 16.43 from the null hypothesis, σ_{x} = 0.8, and n = 15.) You can find the critical value for α = 0.05 in the normal table (see Tables in Appendix B). The z-score for an area to the left equal to 0.05 is midway between –1.65 and –1.64 (0.05 is midway between 0.0505 and 0.0495). The z-score is –1.645. Since –1.645 > –2.08 (which demonstrates that α > the p-value), reject H_0 . Traditionally, the decision to reject or not reject was done in this way. Today, comparing the two probabilities α and the p-value is very common. For this problem, the p-value, 0.0187 is considerably smaller than α, 0.05. You can be confident about your decision to reject. The graph shows α, the p-value, the test statistic, and the critical value.



Note:
Try It
Exercise:

Problem:

The mean throwing distance of a football for a Marco, a high school freshman quarterback, is 40 yards, with a standard deviation of two yards. The team coach tells Marco to adjust his grip to get more distance. The coach records the distances for 20 throws. For the 20 throws, Marco's mean distance was 45 yards. The coach thought the different grip helped Marco throw farther than 40 yards. Conduct a hypothesis test using a preset $\alpha = 0.05$. Assume the throw distances for footballs are normal.

First, determine what type of test this is, set up the hypothesis test, find the *p*-value, sketch the graph, and state your conclusion.

Solution:

Since the problem is about a mean, this is a test of a single population mean.

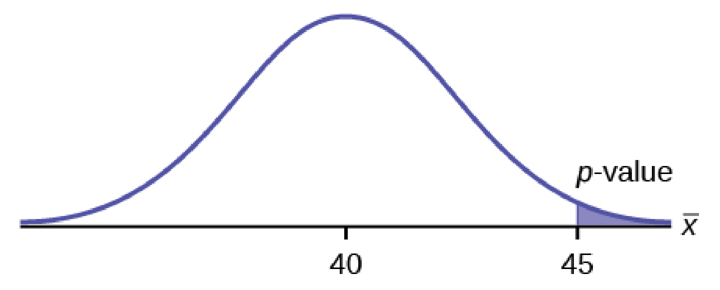
$H_0 : \mu = 40$

$H_a : \mu > 40$

Note:

Press STAT and arrow over to TESTS. Press 1:Z-Test. Arrow over to Stats and press ENTER. Arrow down and enter 40 for μ_0 (null hypothesis), 2 for σ, 45 for the sample mean, and 20 for n. Arrow down to μ: (alternative hypothesis) and set it either as <, ≠, or >. Press ENTER. Arrow down to Calculate and press ENTER. The calculator not only calculates the p-value but it also calculates the test statistic (z-score) for the sample mean. Select <, ≠, or > for the alternative hypothesis. Do this set of instructions again except arrow to Draw (instead of Calculate). Press ENTER. A shaded graph appears with test statistic and p-value. Make sure when you use Draw that no other equations are highlighted in Y = and the plots are turned off.

p-value = $2.6115 \times 10^{-29} \approx 0$



Note: The shading in the figure above is obviously not to scale since an area shaded to scale representing a p-value this small (nearly zero) would not be visible. When you select Draw on your calculator, it will appear as if nothing is shaded, which is a more accurate representation of the p-value.

Because p-value < α, we reject the null hypothesis. There is sufficient evidence to suggest that the change in grip improved Marco’s throwing distance.

Example:
Exercise:

Problem:

A college football coach thought that his players could bench press a **mean weight of 275 pounds**. It is known that the **standard deviation is 55 pounds**. Three of his players thought that the mean weight was **more than** that amount. They asked **30** of their teammates for their estimated maximum lift on the bench press exercise. The data ranged from 205 pounds to 385 pounds. The actual different weights were (frequencies are in parentheses) 205(3) 215(3) 225(1) 241(2) 252(2) 265(2) 275(2) 313(2) 316(5) 338(2) 341(1) 345(2) 368(2) 385(1).

Conduct a hypothesis test using a 2.5% level of significance to determine if the bench press mean is **more than 275 pounds**.

Solution:

Set up the Hypothesis Test:

Since the problem is about a mean weight, this is a **test of a single population mean**.

$H_0 : \mu = 275$
$H_a : \mu > 275$

This is a right-tailed test.

Calculating the distribution needed:

Random variable: $\bar{X}$ = the mean weight, in pounds, lifted by a football player, in a sample of 30 football players.

Distribution for the test: It is normal because σ is known.

$$X \sim N\left(275, \frac{55}{\sqrt{30}}\right)$$

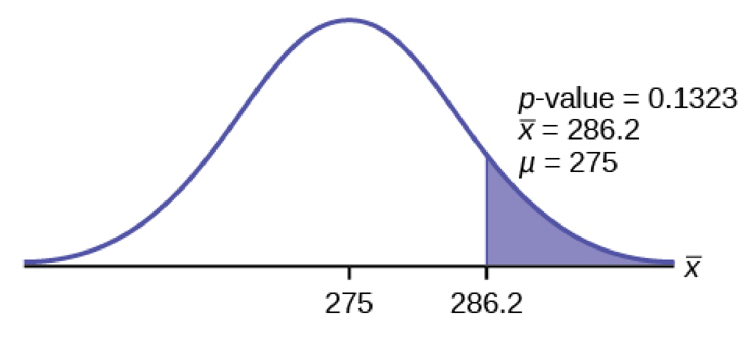
$x = 286.2$ pounds (from the data).

σ = 55 pounds **(Always use σ if you know it.)** We assume μ = 275 pounds to run the test.

Calculate the p-value using the normal distribution for a mean and using the sample mean as input (see [link] for using the data as input):

$$p\text{-value} = P(x > 286.2) = 0.1323.$$

Interpretation of the *p*-value: If H_0 is true, then there is a 0.1323 probability (13.23%) that 30 randomly chosen football players can lift a mean weight of 286.2 pounds or more. Because a 13.23% chance is large enough, a mean weight lift of 286.2 pounds or more is not a rare event.



Compare α and the p-value:

$\alpha = 0.025$ and the p-value = 0.1323

Make a decision: Since p-value > α, do not reject H_0 .

Conclusion: At the 2.5% level of significance, from the sample data, there is not sufficient evidence to conclude that the true mean weight lifted is more than 275 pounds.

The p-value can easily be calculated using the following steps:

Note:
Put the data and frequencies into lists. Press `STAT` and arrow over to `TESTS`. Press `1:Z-Test`. Arrow over to `Data` and press `ENTER`. Arrow down and enter 275 for μ_0, 55 for σ, the name of the list where you put the data, and the name of the list where you put the frequencies. Arrow down to μ and arrow over to > μ_0, since $\mu > 275$ is the alternative hypothesis. Press `ENTER`. Arrow down to `Calculate` and press `ENTER`. The calculator not only calculates the p-value ($p = 0.1331$, a little different from the previous calculation, in which we used the sample mean rounded to one decimal place instead of the actual data), but it also calculates the test statistic (z-score) for the sample mean, the sample mean, and the sample standard deviation. Do this set of instructions again except arrow to `Draw` (instead of `Calculate`). Press `ENTER`. A shaded graph appears with $z = 1.112$ (test statistic) and $p = 0.1331$ (p-value). Make sure when you use `Draw` that no other equations are highlighted in Y = and the plots are turned off.

Example:
Exercise:

Problem:

Statistics students believe that the mean score on the first statistics test is 65. A statistics instructor thinks the mean score is higher than 65. He samples ten statistics students and obtains the scores 65 65 70 67 66 63 63 68 72 71. He performs a hypothesis test using a 5% level of significance. The data are assumed to be from a normal distribution.

Solution:

Set up the hypothesis test:

A 5% level of significance means that $\alpha = 0.05$. This is a test of a **single population mean**.

$H_0 : \mu = 65 \qquad H_a : \mu > 65$

Since the instructor thinks the average score is higher, use a ">" for the alternative hypothesis. The ">" means the test is right-tailed.

Determine the distribution needed:

Random variable: X = average score on the first statistics test, in a sample of 10 statistics students.

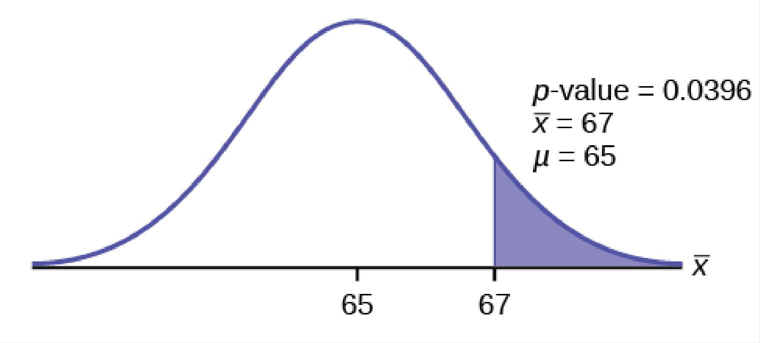
Distribution for the test: If you read the problem carefully, you will notice that there is **no population standard deviation given**. You are only given $n = 10$ sample data values. Notice also that the data come from a normal distribution. This means that the distribution for the test is a student's t.

Use t_{df}. Therefore, the distribution for the test is t_9 where n = 10 and *df* = 10 - 1 = 9.

Calculate the p-value using the Student's t-distribution:

p-value = $P(x > 67)$ = 0.0396 where the sample mean and sample standard deviation are calculated as 67 and 3.1972 from the data.

Interpretation of the p-value: If the null hypothesis is true, then there is a 0.0396 probability (3.96%) that a sample mean will be 67 or more.



Compare α and the p-value:

Since α = 0.05 and the p-value = 0.0396, the p-value < α.

Make a decision: Since the p-value < α, reject H_0 .

This means you reject μ = 65. In other words, you believe the average test score is more than 65.

Conclusion: At a 5% level of significance, the sample data show sufficient evidence that the mean (average) test score is more than 65, just as the math instructor thinks.

The p-value can easily be calculated as follows:

Note:
Put the data into a list. Press `STAT` and arrow over to `TESTS`. Press `2:T-Test`. Arrow over to `Data` and press `ENTER`. Arrow down and enter 65 for μ_0, the name of the list where you put the data, and 1 for `Freq:`. Arrow down to μ: and arrow over to > μ_0, since $\mu > 65$ is the alternative hypothesis.. Press `ENTER`. Arrow down to `Calculate` and press `ENTER`. The calculator not only calculates the p-value (p = 0.0396) but it also calculates the test statistic (t-score) for the sample mean, the sample mean, and the sample standard deviation. Do this set of instructions again except arrow to `Draw` (instead of `Calculate`). Press `ENTER`. A shaded graph appears with t = 1.9781 (test statistic) and p = 0.0396 (p-value). Make sure when you use `Draw` that no other equations are highlighted in Y = and the plots are turned off.

Note:
Try It
Exercise:

Problem:

It is believed that a stock price for a particular company will grow at a rate of \$5 per week with a standard deviation of \$1. An investor believes the stock won't grow as quickly. The changes in stock price is recorded for ten weeks and are as follows: \$4, \$3, \$2, \$3, \$1, \$7, \$2, \$1, \$1, \$2. Perform a hypothesis test using a 5% level of significance. State the null and alternative hypotheses, find the *p*-value and state your conclusion.

Solution:

$H_0 : \mu = 5$

$H_a : \mu < 5$

p-value = 0.0082

Because the *p*-value < α, we reject the null hypothesis. There is sufficient evidence to suggest that the stock price of the company grows at a rate less than \$5 a week.

Example:
Exercise:

Problem:

Joon believes that 50% of first-time brides in the United States are younger than their grooms. She performs a hypothesis test to determine if the percentage is **the same or different from 50%**. Joon randomly samples **100 first-time brides** and **53** reply that they are younger than their grooms. For the hypothesis test, she uses a 1% level of significance.

Solution:

Set up the hypothesis test:

The 1% level of significance means that $\alpha = 0.01$. This is a **test of a single population proportion**.

$H_0 : p = 0.50$ $H_a : p \neq 0.50$

The words **"is the same or different from"** tell you this is a two-tailed test.

Calculate the distribution needed:

Random variable: $\widehat{P}$ = the percent of first-time brides who are younger than their grooms, in a sample of 100.

Distribution for the test: The problem contains no mention of a mean. The information is given in terms of percentages. Use the distribution for $\widehat{P}$, the estimated proportion.

$\widehat{P} \sim N\left(p, \sqrt{\frac{p \cdot q}{n}}\right)$, where $p = 0.50$, $q = 1 - p = 0.50$, and $n = 100$

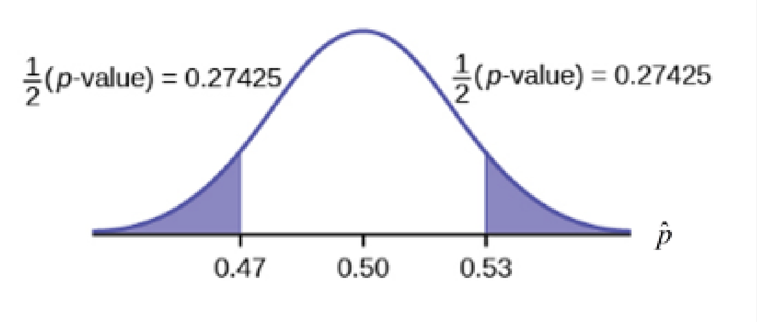
Therefore, $\widehat{P} \sim N\left(0.5, \sqrt{\frac{0.5 \cdot 0.5}{100}}\right)$

Calculate the p-value using the normal distribution for proportions:

p-value = $P(\widehat{p} < 0.47 \text{ or } \widehat{p} > 0.53) = 0.5485$

where $x = 53$, $\widehat{p} = \frac{x}{n} = \frac{53}{100} = 0.53$.

Interpretation of the p-value: If the null hypothesis is true, there is 0.5485 probability (54.85%) that a sample (estimated) proportion $\widehat{p}$ would be 0.53 or more OR 0.47 or less (see graph below).



The mean of this distribution, $\mu = p = 0.50$ comes from H_0 , the null hypothesis.

The sample proportion, $\widehat{p} = 0.53$. Since the curve is symmetrical and the test is two-tailed, the $\widehat{p}$ for the left tail is equal to 0.50 – 0.03 = 0.47 where $\mu = p = 0.50$. (0.03 is the difference between 0.53 and 0.50.)

Compare α and the p-value:

Since $\alpha = 0.01$ and the p-value = 0.5485, the p-value > α.

Make a decision: Since p-value > α, you cannot reject H_0 .

Conclusion: At the 1% level of significance, the sample data do not show sufficient evidence that the percentage of first-time brides who are younger than their grooms is different from 50%.

The p-value can easily be calculated as follows:

Note:
Press `STAT` and arrow over to `TESTS`. Press `5:1-PropZTest`. Enter 0.5 for p_0, 53 for x, and 100 for n. Arrow down to `Prop` and arrow to `≠` p_0, since `Prop ≠` 0.5 is the alternate hypothesis. Press `ENTER`. Arrow down to `Calculate` and press `ENTER`. The calculator calculates the p-value (p = 0.5485) and the test statistic (z-score). Do this set of instructions again except arrow to `Draw` (instead of `Calculate`). Press `ENTER`. A shaded graph appears with z = 0.6 (test statistic) and p = 0.5485 (p-value). Make sure when you use `Draw` that no other equations are highlighted in Y = and the plots are turned off.

Note:
Try It
Exercise:

Problem:

A teacher believes that 85% of students in the class will want to go on a field trip to the local zoo. She performs a hypothesis test to determine if the percentage is the same or different from 85%. The teacher randomly samples 50 students and 39 reply that they would want to go to the zoo. For the hypothesis test, use a 1% level of significance.

First, determine what type of test this is. Then, set up the hypothesis test, find the *p*-value, sketch the graph, and state your conclusion.

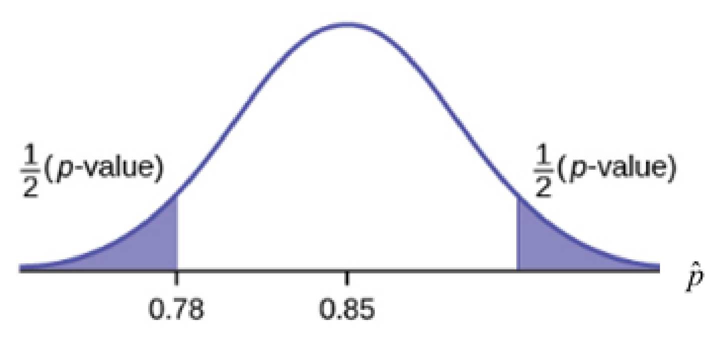
Solution:

Since the problem is about percentages, this is a test of a single population proportion.

$H_0 : p = 0.85$

$H_a : p \neq 0.85$

p-value = 0.1657



Because the *p*-value > α, we fail to reject the null hypothesis. There is not sufficient evidence to suggest that the proportion of students that want to go to the zoo is not 85%.

Example:
Exercise:

Problem:

Suppose a consumer group suspects that the proportion of households that have three cell phones is 30%. A cell phone company has reason to believe that the proportion is not 30%. Before they start a big advertising campaign, they conduct a hypothesis test. Their marketing people randomly survey 150 households with the result that 43 of the households have three cell phones.

Solution:

Set up the Hypothesis Test:

$H_0 : p = 0.30$; $H_a : p \neq 0.30$

Determine the distribution needed:

The **random variable** is $\widehat{P}$ = proportion of households that have three cell phones, in a sample of 150 households.

The **distribution** for the hypothesis test is $\widehat{P} \sim N\left(0.3, \sqrt{\frac{(0.3)(0.7)}{150}}\right)$

Exercise:

Problem:

a. The value that helps determine the p-value is $\widehat{p}$. Calculate $\widehat{p}$.

Solution:

a. $\hat{p}$ where x is the number of successes and n is the total number in the sample.

x = 43, n = 150

$\hat{p} = \frac{43}{150}$

Exercise:

Problem: b. What is a **success** for this problem?

Solution:

b. A success is having three cell phones in a household.

Exercise:

Problem: c. What is the level of significance?

Solution:

c. The level of significance is the preset α. Since α is not given, assume that α = 0.05.

Exercise:

Problem:

d. Draw the graph for this problem. Draw the horizontal axis. Label and shade appropriately.
Calculate the p-value.

Solution:

d. p-value = 0.7216

Exercise:

Problem:

e. Make a decision. ______________(Reject/Do not reject) H_0 because_____________.

Solution:

e. Assuming that $\alpha = 0.05$, p-value > α. The decision is do not reject H_0 because there is not sufficient evidence to conclude that the proportion of households that have three cell phones is not 30%.

Note:
Try It
Exercise:

Problem:

Marketers believe that 92% of adults in the United States own a cell phone. A cell phone manufacturer believes that number is actually lower. 200 American adults are surveyed, of which, 174 report having cell phones. Using a 5% level of significance, is there evidence to support the manufacturer's belief? State the null and alternative hypotheses, find the p-value, and state your conclusion.

Solution:

$H_0 : p = 0.92$

$H_a : p < 0.92$

p-value = 0.0046

Because $p < 0.05$, we reject the null hypothesis. There is sufficient evidence to conclude that fewer than 92% of American adults own cell phones.

The next example is a poem written by a statistics student named Nicole Hart. The solution to the problem follows the poem. Notice that the hypothesis test is for a single population proportion. This means that the null and alternate hypotheses use the parameter p. The distribution for the test is normal. The estimated proportion $\widehat{p}$ is the proportion of fleas killed to the total fleas found on Fido. This is sample information. The problem gives a preconceived $\alpha = 0.01$, for comparison, and a 95% confidence interval computation. The poem is clever and humorous, so please enjoy it!

Example:
Exercise:

My dog has so many fleas,
They do not come off with ease.
As for shampoo, I have tried many types
Even one called Bubble Hype,
Which only killed 25% of the fleas,
Unfortunately I was not pleased.

I've used all kinds of soap,
Until I had given up hope
Until one day I saw
An ad that put me in awe.

A shampoo used for dogs
Called GOOD ENOUGH to Clean a Hog
Guaranteed to kill more fleas.

I gave Fido a bath
And after doing the math
His number of fleas
Started dropping by 3's!

Before his shampoo
I counted 42.
At the end of his bath,
I redid the math
And the new shampoo had killed 17 fleas.
So now I was pleased.

Now it is time for you to have some fun
With the level of significance being .01,
You must help me figure out
Problem: Use the new shampoo or go without?

Solution:

Set up the hypothesis test:

$H_0 : p = 0.25$ $H_a : p > 0.25$

Determine the distribution needed:

In words, CLEARLY state what your random variable X or $\widehat{P}$ represents.

$\widehat{P}$ = The proportion of fleas that are killed by the new shampoo

State the distribution to use for the test.

Normal: $N\left(0.25, \sqrt{\frac{(0.25)(1-0.25)}{42}}\right)$

Test Statistic: $z = 2.3163$

Calculate the p-value using the normal distribution for proportions:

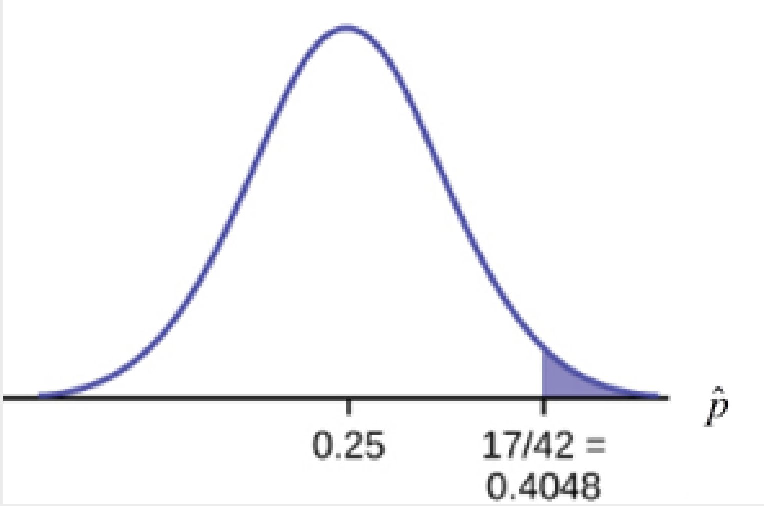
p-value = 0.0103

In one to two complete sentences, explain what the p-value means for this problem.

If the null hypothesis is true (the proportion is 0.25), then there is a 0.0103 probability that the sample (estimated) proportion is 0.4048

$\left(\frac{17}{42}\right)$ or more.

Use the previous information to sketch a picture of this situation. CLEARLY, label and scale the horizontal axis and shade the region(s) corresponding to the p-value.



Compare α and the p-value:

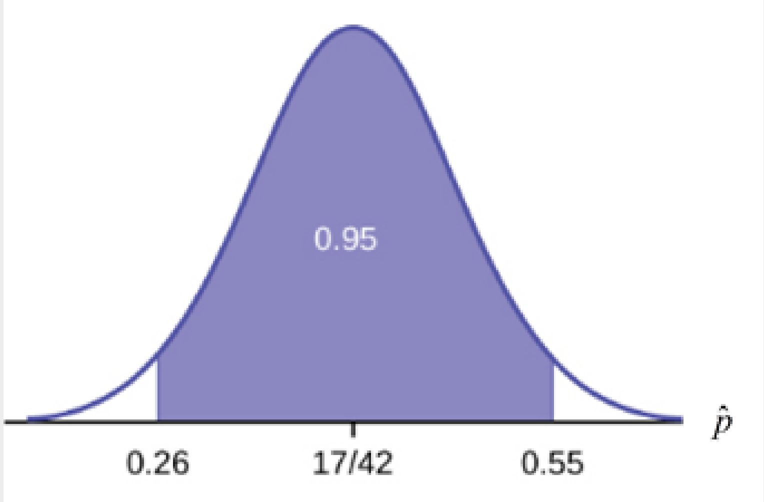
Since α = 0.01 and the p-value = 0.0103, the p-value > α.

Indicate the correct decision ("reject" or "do not reject" the null hypothesis), the reason for it, and write an appropriate conclusion, using complete sentences.

alpha	decision	reason for decision
0.01	Do not reject H_0	p-value > α

Conclusion: At the 1% level of significance, the sample data do not show sufficient evidence that the percentage of fleas that are killed by the new shampoo is more than 25%.

Construct a 95% confidence interval for the true mean or proportion. Include a sketch of the graph of the situation. Label the point estimate and the lower and upper bounds of the confidence interval.



Confidence Interval: (0.26,0.55) We are 95% confident that the true population proportion *p* of fleas that are killed by the new shampoo is between 26% and 55%.

Note:
Note
This test result is not very definitive since the *p*-value is very close to alpha. In reality, one would probably do more tests by giving the dog another bath after the fleas have had a chance to return.

Example:
Exercise:

Problem:

The National Institute of Standards and Technology provides exact data on conductivity properties of materials. Following are conductivity measurements for 11 randomly selected pieces of a particular type of glass.

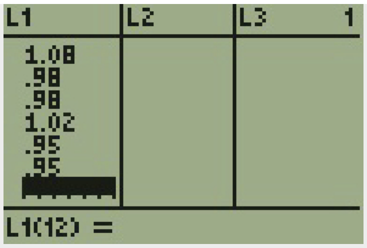
1.11; 1.07; 1.11; 1.07; 1.12; 1.08; .98; .98 1.02; .95; .95

Is there convincing evidence that the average conductivity of this type of glass is greater than one? Use a significance level of 0.05. Assume the population is normal.

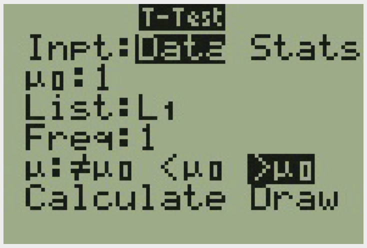
Solution:

Let’s follow a four-step process to answer this statistical question.

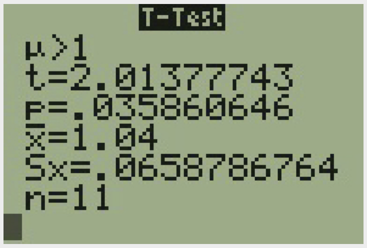
1. **State the Question**: We need to determine if, at a 0.05 significance level, the average conductivity of the selected glass is greater than one. Our hypotheses will be
    a. $H_0 : \mu = 1$
    b. $H_a : \mu > 1$
2. **Plan**: We are testing a sample mean without a known population standard deviation. Therefore, we need to use a Student's t-distribution. Assume the underlying population is normal.
3. **Do the calculations**: We will input the sample data into the TI-83 as follows.



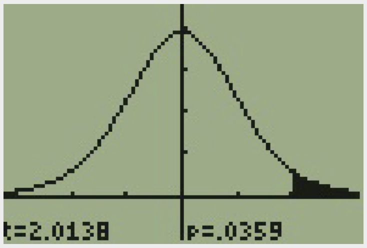
```
L1     L2     L3     1
1.08
.98
.98
1.02
.95
.95
L1(12) =
```



```
T-Test
Inpt:Data Stats
μ0:1
List:L1
Freq:1
μ:≠μ0 <μ0 >μ0
Calculate Draw
```



```
T-Test
μ>1
t=2.01377743
p=.035860646
x̄=1.04
Sx=.0658786764
n=11
```



4. **State the Conclusions**: Since the *p*-value ($p = 0.036$) is less than our alpha value, we will reject the null hypothesis. It is reasonable to state that the data supports the claim that the average conductivity level is greater than one.

Example:
Exercise:

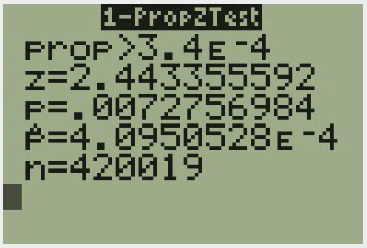
Problem:

In a study of 420,019 cell phone users, 172 of the subjects developed brain cancer. Test the claim that cell phone users developed brain cancer at a greater rate than that for non-cell phone users (the rate of brain cancer for non-cell phone users is 0.0340%). Since this is a critical issue, use a 0.005 significance level.

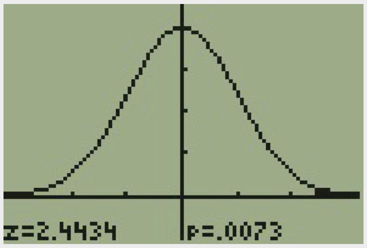
Solution:

We will follow the four-step process.

1. We need to conduct a hypothesis test on the claimed cancer rate. Our hypotheses will be
    a. $H_0 : p = 0.00034$
    b. $H_a : p > 0.00034$
2. We will be testing a sample proportion with $x = 172$ and $n = 420{,}019$. The sample is sufficiently large because we have $np = 420{,}019(0.00034) = 142.8$, $nq = 420{,}019(0.99966) = 419{,}876.2$, two independent outcomes, and a fixed probability of success $p = 0.00034$. Thus we will be able to generalize our results to the population.
3. The associated TI results are



```
1-PropZTest
prop>3.4E-4
z=2.443355592
p=.0072756984
p̂=4.0950528E-4
n=420019
```



4. Since the p-value = 0.0073 is greater than our alpha value = 0.005, we cannot reject the null. Therefore, we conclude that there is not enough evidence to support the claim of higher brain cancer rates for the cell phone users.

Example:
Exercise:

Problem:

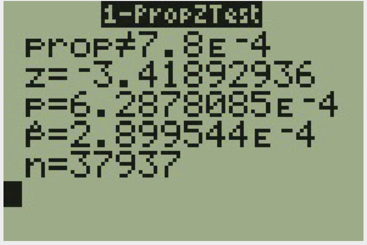
According to the US Census there are approximately 268,608,618 residents aged 12 and older. Statistics from the Rape, Abuse, and Incest National Network indicate that, on average, 207,754 rapes occur each year (male and female) for persons aged 12 and older. This translates into a percentage of sexual assaults of 0.078%. In Daviess County, KY, there were reported 11 rapes for a population of 37,937. Conduct an appropriate hypothesis test to determine if there is a statistically significant difference between the local sexual assault percentage and the national sexual assault percentage. Use a significance level of 0.01.

Solution:

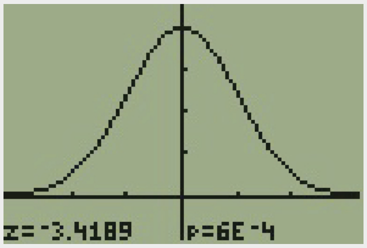
We will follow the four-step process.

1. We need to test whether the proportion of sexual assaults in Daviess County, KY is significantly different from the national average.
2. Since we are presented with proportions, we will use a one-proportion z-test. The hypotheses for the test will be
   a. $H_0 : p = 0.00078$
   b. $H_a : p \neq 0.00078$

3. The following screen shots display the summary statistics from the hypothesis test.



```
1-PropZTest
prop≠7.8E-4
z=-3.41892936
p=6.2878085E-4
p̂=2.899544E-4
n=37937
```



4. Since the p-value, $p = 0.00063$, is less than the alpha level of 0.01, the sample data indicates that we should reject the null hypothesis. In conclusion, the sample data support the claim that the proportion of sexual assaults in Daviess County, Kentucky is different from the national proportion.

Section Review

The **hypothesis test** itself has an established process. This can be summarized as follows:

1. Determine H_0 and H_a.
2. Determine the random variable.
3. Determine the distribution for the test.
4. Draw a graph, calculate the test statistic, and use the test statistic to calculate the p-value. (A z-score and a t-score are examples of test statistics.)
5. Compare the preconceived α with the p-value, make a decision (reject or do not reject H_0), and write a clear conclusion using English sentences.

Exercise:

Problem:

Assume $H_0 : \mu = 9$ and $H_a : \mu < 9$. Is this a left-tailed, right-tailed, or two-tailed test?

Solution:

This is a left-tailed test.

Exercise:

Problem:

Assume $H_0 : \mu = 6$ and $H_a : \mu > 6$. Is this a left-tailed, right-tailed, or two-tailed test?

Exercise:

Problem:

Assume $H_0 : p = 0.25$ and $H_a : p \neq 0.25$. Is this a left-tailed, right-tailed, or two-tailed test?

Solution:

This is a two-tailed test.

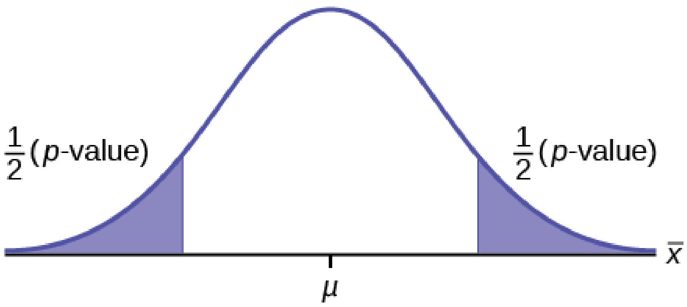
Exercise:

Problem: Draw the general graph of a left-tailed test.

Exercise:

Problem: Draw the general graph of a two-tailed test.

Solution:



Exercise:

Problem:

A bottle of water is labeled as containing 16 fluid ounces of water. You believe it is less than that. What type of test would you use?

Exercise:

Problem:

Your friend claims that his mean golf score is 63. You want to show that it is higher than that. What type of test would you use (right-, left-, or two-tailed)?

Solution:

a right-tailed test

Exercise:

Problem:

A bathroom scale claims to be able to identify correctly any weight within a pound. You think that it cannot be that accurate. What type of test would you use (right-, left-, or two-tailed)?

Exercise:

Problem:

You flip a coin and record whether it shows heads or tails. You know the probability of getting heads is 50%, but you think it is less for this particular coin. What type of test would you use (right-, left-, or two-tailed)?

Solution:

a left-tailed test

Exercise:

Problem:

If the alternative hypothesis has a not equals (≠) symbol, you know to use which type of test (right-, left-, or two-tailed)?

Exercise:

Problem:

Assume the alternative hypothesis states that the mean is less than 18. Is this a left-tailed, right-tailed, or two-tailed test?

Solution:

This is a left-tailed test.

Exercise:

Problem:

Assume the alternative hypothesis states that the mean is more than 12. Is this a left-tailed, right-tailed, or two-tailed test?

Exercise:

Problem:

Assume the null hypothesis states that the mean is equal to 88. The alternative hypothesis states that the mean is not equal to 88. Is this a left-tailed, right-tailed, or two-tailed test?

Solution:

This is a two-tailed test.

Homework

For each of the word problems below, use a solution sheet to do the hypothesis test. The Solutions Sheets can be found in the Table of Contents or by clicking here. Please feel free to make copies of the Solution Sheets. For the online version of the book, it is suggested that you copy the .doc or the .pdf files.

Note:
Note
If you are using a Student's t-distribution for one of the following homework problems, you may assume that the underlying population is normally distributed. (In general, you must first prove that assumption, however.)

Exercise:

Problem:

A particular brand of tires claims that its deluxe tire averages at least 50,000 miles before it needs to be replaced. From past studies of this tire, the standard deviation is known to be 8,000. A survey of owners of that tire design is conducted. From the 28 tires surveyed, the mean lifespan was 46,500 miles with a standard deviation of 9,800 miles. Using alpha = 0.05, is the data highly inconsistent with the claim? In other words, is there convincing evidence (at the 5% significance level) that the deluxe tires actually average less than 50,000 miles before needing to be replaced?

Solution:

a. H_0 : μ = 50,000
b. H_a : μ < 50,000
c. Let $\bar{X}$ = the average lifespan of a particular brand of tire, in a sample of 28 tires.
d. normal distribution
e. z = -2.315
f. p-value = 0.0103
g. Check student's solution.

h.
  i. alpha: 0.05
  ii. Decision: Reject the null hypothesis.
  iii. Reason for decision: The p-value is less than 0.05.
  iv. Conclusion: There is sufficient evidence to conclude that the mean lifespan of the tires is less than 50,000 miles.

i. 95% confidence interval: (43,537, 49,463)

Exercise:

Problem:

From generation to generation, the mean age when smokers first start to smoke varies. However, the standard deviation of that age remains constant at around 2.1 years. A survey of 40 smokers of this generation was done to see if the mean starting age is at least 19. The sample mean was 18.1 with a sample standard deviation of 1.3. Do the data support the claim at the 5% level?

Exercise:

Problem:

The cost of a daily newspaper varies from city to city. However, the variation among prices remains steady with a standard deviation of 20¢. A study was done to test the claim that the mean cost of a daily newspaper is $1.00. Twelve costs yield a mean cost of 95¢ with a standard deviation of 18¢. Do the data support the claim at the 1% level? Assume the cost of a daily newspaper is normally distributed.

Solution:

a. $H_0 : \mu = \$1.00$
b. $H_a : \mu \neq \$1.00$
c. Let X = the average cost of a daily newspaper, in a sample of 12 newspapers.
d. normal distribution
e. $z = –0.866$
f. p-value = 0.3865
g. Check student’s solution.

h.
  i. Alpha: 0.01
  ii. Decision: Do not reject the null hypothesis.
  iii. Reason for decision: The p-value is greater than 0.01.
  iv. Conclusion: There is not sufficient evidence to reject the claim that the mean cost of daily papers is $1. The mean cost could be $1.

i. 95% confidence interval: ($0.84, $1.06)

Exercise:

Problem:

An article in the *San Jose Mercury News* stated that students in the California state university system take 4.5 years, on average, to finish their undergraduate degrees. Suppose you believe that the mean time is longer. You conduct a survey of 49 students and obtain a sample mean of 5.1 with a sample standard deviation of 1.2. Do the data support your claim at the 1% level?

Exercise:

Problem:

The mean number of sick days an employee takes per year is believed to be about ten. Members of a personnel department do not believe this figure. They randomly survey eight employees. The number of sick days they took for the past year are as follows: 12; 4; 15; 3; 11; 8; 6; 8. Should the personnel team believe that the mean number is ten?

Solution:

a. $H_0 : \mu = 10$
b. $H_a : \mu \neq 10$
c. Let $\bar{X}$ the mean number of sick days an employee takes per year, in a sample of 8 employees.
d. Student's t-distribution
e. $t = –1.12$
f. p-value = 0.300
g. Check student's solution.

h.
  i. Alpha: 0.05
  ii. Decision: Do not reject the null hypothesis.
  iii. Reason for decision: The p-value is greater than 0.05.
  iv. Conclusion: At the 5% significance level, there is insufficient evidence to conclude that the mean number of

sick days is not ten.

i. 95% confidence interval: (4.9443, 11.806)

Exercise:

Problem:

In 1955, *Life Magazine* reported that the 25 year-old mother of three worked, on average, an 80 hour week. Recently, many groups have been studying whether or not the women's movement has, in fact, resulted in an increase in the average work week for women (combining employment and at-home work). Suppose a study was done to determine if the mean work week has increased. 81 women were surveyed with the following results. The sample mean was 83; the sample standard deviation was ten. Does it appear that the mean work week has increased for women at the 5% level?

Exercise:

Problem:

Your statistics instructor claims that 60 percent of the students who take her Elementary Statistics class go through life feeling more enriched. For some reason that she can't quite figure out, most people don't believe her. You decide to check this out on your own. You randomly survey 64 of her past Elementary Statistics students and find that 34 feel more enriched as a result of her class. Now, what do you think?

Solution:

a. $H_0 : p = 0.6$
b. $H_a : p < 0.6$
c. Let $\widehat{P}$ = the proportion of students who feel more enriched as a result of taking Elementary Statistics, in a sample of 64 students.
d. normal for a single proportion
e. z = -1.12
f. p-value = 0.1308

g. Check student's solution.

h.
  i. Alpha: 0.05
  ii. Decision: Do not reject the null hypothesis.
  iii. Reason for decision: The *p*-value is greater than 0.05.
  iv. Conclusion: There is insufficient evidence to conclude that less than 60 percent of her students feel more enriched.

i. 95% Confidence Interval: (0.409, 0.654)
The "plus-4" 95% confidence interval is (0.411, 0.648)

Exercise:

Problem:

According to an article in *Newsweek*, the natural ratio of girls to boys is 100 to 105. In China, the birth ratio is 100 to 114 (46.7% girls). Suppose you don't believe the reported figures of the percent of girls born in China. You conduct a study. In this study, you count the number of girls and boys born in 150 randomly chosen recent births. There are 60 girls and 90 boys born of the 150. Based on your study, do you believe that the percent of girls born in China is 46.7?

Exercise:

Problem:

A poll done for *Newsweek* found that 13% of Americans have seen or sensed the presence of an angel. A contingent doubts that the percent is really that high. It conducts its own survey. Out of 76 Americans randomly surveyed, only two had seen or sensed the presence of an angel. As a result of the contingent's survey, would you agree with the *Newsweek* poll? In complete sentences, also give three reasons why the two polls might give different results.

Solution:

a. H_0: $p = 0.13$
b. H_a: $p < 0.13$

c. Let $\widehat{P}$ = the proportion of Americans who have seen or sensed angels, in a sample of 76 Americans.
d. normal for a single proportion
e. $z = –2.688$
f. p-value = 0.0036
g. Check student's solution.

h.
  i. alpha: 0.05
  ii. Decision: Reject the null hypothesis.
  iii. Reason for decision: The p-value is less than 0.05.
  iv. Conclusion: There is sufficient evidence to conclude that the percentage of Americans who have seen or sensed an angel is less than 13%.

i. 95% confidence interval: (-0.0097, 0.0623).
The "plus-4" 95% confidence interval is (0.0022, 0.0978)

Exercise:

Problem:

The mean work week for engineers in a start-up company is believed to be about 60 hours. A newly hired engineer hopes that it's shorter. She asks ten engineering friends in start-ups for the lengths of their mean work weeks. Based on the results that follow, should she count on the mean work week to be shorter than 60 hours?

Data (length of mean work week): 70; 45; 55; 60; 65; 55; 55; 60; 50; 55.

Exercise:

Problem:

Use the "Lap time" data for Lap 4 (see [link]) to test the claim that Terri finishes Lap 4, on average, in less than 129 seconds. Use all twenty races given.

Solution:

a. $H_0 : \mu = 129$
b. $H_a : \mu < 129$
c. Let $\bar{X}$ = the average time in seconds that Terri finishes Lap 4, in a sample of 20 races.
d. Student's t-distribution
e. $t = 1.209$
f. p-value = 0.8792
g. Check student’s solution.

h.
   i. Alpha: 0.05
   ii. Decision: Do not reject the null hypothesis.
   iii. Reason for decision: The p-value is greater than 0.05.
   iv. Conclusion: There is insufficient evidence to conclude that Terri’s mean lap time is less than 129 seconds.

i. 95% confidence interval: (128.63, 130.37)

Exercise:

Problem:

Use the “Initial Public Offering” data (see [link]) to test the claim that the mean offer price was $18 per share. Do not use all the data. Use your random number generator to randomly survey 15 prices.

Exercise:

Problem:

Toastmasters International cites a report by Gallop Poll that 40% of Americans fear public speaking. A student believes that less than 40% of students at her school fear public speaking. She randomly surveys 361 schoolmates and finds that 135 report they fear public speaking. Conduct a hypothesis test to determine if the percent at her school is less than 40%.

Solution:

a. $H_0 : p = 0.40$

b. $H_a : p < 0.40$
c. Let $\widehat{P}$ = the proportion of schoolmates who fear public speaking, in a sample of 361 students.
d. normal for a single proportion
e. $z = –1.01$
f. p-value = 0.1563
g. Check student’s solution.

h.
  i. Alpha: 0.05
  ii. Decision: Do not reject the null hypothesis.
  iii. Reason for decision: The p-value is greater than 0.05.
  iv. Conclusion: There is insufficient evidence to support the claim that less than 40% of students at the school fear public speaking.

i. 95% confidence interval: (0.3241, 0.4239): The 95% “plus-4” confidence interval is (0.3257, 0.4250).

Exercise:

Problem:

Sixty-eight percent of online courses taught at community colleges nationwide were taught by full-time faculty. To test if 68% also represents California’s percent for full-time faculty teaching the online classes, Long Beach City College (LBCC) in California was randomly selected for comparison. In the same year, 34 of the 44 online courses LBCC offered were taught by full-time faculty. Conduct a hypothesis test to determine if 68% represents California.

Exercise:

Problem:

According to an article in *Bloomberg Businessweek*, New York City's most recent adult smoking rate is 14%. Suppose that a survey is conducted to determine this year's rate. Nine out of 70 randomly chosen N.Y. City residents reply that they smoke. Conduct a hypothesis test to determine if the rate is still 14% or if it has decreased.

Solution:

a. $H_0 : p = 0.14$
b. $H_a : p < 0.14$
c. Let $\widehat{P}$ = the proportion of NYC residents that smoke, in a sample of 70 NYC residents.
d. normal for a single proportion
e. $z = –0.2756$
f. p-value = 0.3914
g. Check student's solution.

h.
   i. alpha: 0.05
   ii. Decision: Do not reject the null hypothesis.
   iii. Reason for decision: The p-value is greater than 0.05.
   iv. At the 5% significance level, there is insufficient evidence to conclude that the proportion of NYC residents who smoke is less than 0.14.

i. 95% confidence interval: (0.0502, 0.2070): The 95% "plus-4" confidence interval (see chapter 8) is (0.0676, 0.2297).

Exercise:

Problem:

The mean age of De Anza College students in a previous term was 26.6 years old. An instructor thinks the mean age for online students is older than 26.6. She randomly surveys 56 online students and finds that the sample mean is 29.4 with a standard deviation of 2.1. Conduct a hypothesis test.

Exercise:

Problem:

Registered nurses earned an average annual salary of $69,110. For that same year, a survey was conducted of 41 California registered nurses to determine if the annual salary is higher than $69,110 for California nurses. The sample average was $71,121 with a sample standard deviation of $7,489. Conduct a hypothesis test.

Solution:

a. $H_0 : \mu = 69{,}110$
b. $H_a : \mu > 69{,}110$
c. Let $\bar{X}$ = the mean salary in dollars for a sample of 41 California registered nurses.
d. Student's t-distribution
e. $t = 1.719$
f. p-value: 0.0466
g. Check student’s solution.

h.
    i. Alpha: 0.05
    ii. Decision: Reject the null hypothesis.
    iii. Reason for decision: The p-value is less than 0.05.
    iv. Conclusion: At the 5% significance level, there is sufficient evidence to conclude that the mean salary of California registered nurses exceeds $69,110.

i. 95% confidence interval: ($68,757, $73,485)

Exercise:

Problem:

La Leche League International reports that the mean age of weaning a child from breastfeeding is age four to five worldwide. In America, most nursing mothers wean their children much earlier. Suppose a random survey is conducted of 21 U.S. mothers who recently weaned their children. The mean weaning age was nine months (3/4 year) with a standard deviation of 4 months. Conduct a hypothesis test to determine if the mean weaning age in the U.S. is less than four years old.

Exercise:

Problem:

Over the past few decades, public health officials have examined the link between weight concerns and teen girls' smoking. Researchers surveyed a group of 273 randomly selected teen girls living in Massachusetts (between 12 and 15 years old). After four years the girls were surveyed again. Sixty-three said they smoked to stay thin. Is there good evidence that more than thirty percent of the teen girls smoke to stay thin?
After conducting the test, your decision and conclusion are

a. Reject H_0 : There is sufficient evidence to conclude that more than 30% of teen girls smoke to stay thin.
b. Do not reject H_0 : There is not sufficient evidence to conclude that less than 30% of teen girls smoke to stay thin.
c. Do not reject H_0 : There is not sufficient evidence to conclude that more than 30% of teen girls smoke to stay thin.
d. Reject H_0 : There is sufficient evidence to conclude that less than 30% of teen girls smoke to stay thin.

Solution:

c

Exercise:

Problem:

A statistics instructor believes that fewer than 20% of Evergreen Valley College (EVC) students attended the opening night midnight showing of the latest Harry Potter movie. She surveys 84 of her students and finds that 11 of them attended the midnight showing. At a 1% level of significance, an appropriate conclusion is:

a. There is insufficient evidence to conclude that the percent of EVC students who attended the midnight showing of Harry Potter is less than 20%.
b. There is sufficient evidence to conclude that the percent of EVC students who attended the midnight showing of Harry Potter is more than 20%.
c. There is sufficient evidence to conclude that the percent of EVC students who attended the midnight showing of Harry Potter is less than 20%.
d. There is insufficient evidence to conclude that the percent of EVC students who attended the midnight showing of Harry Potter is at least 20%.

Exercise:

Problem:

Previously, an organization reported that teenagers spent 4.5 hours per week, on average, on the phone. The organization thinks that, currently, the mean is higher. Fifteen randomly chosen teenagers were asked how many hours per week they spend on the phone. The sample mean was 4.75 hours with a sample standard deviation of 2.0. Conduct a hypothesis test.

At a significance level of alpha = 0.05, what is the correct conclusion?

a. There is enough evidence to conclude that the mean number of hours is more than 4.75

b. There is enough evidence to conclude that the mean number of hours is more than 4.5
c. There is not enough evidence to conclude that the mean number of hours is more than 4.5
d. There is not enough evidence to conclude that the mean number of hours is more than 4.75

Solution:

c

Instructions: For the following ten exercises, answer each of the questions listed below.

a. What is the null and the alternate hypothesis?
b. What is the p-value?
c. What is alpha?
d. What is your decision?
e. What is your conclusion?
f. Answer any other questions asked in the problem.

Exercise:

Problem:

According to the Center for Disease Control website, in 2011 at least 18% of high school students have smoked a cigarette. An Introduction to Statistics class in Davies County, KY conducted a hypothesis test at the local high school (a medium sized–approximately 1,200 students–small city demographic) to determine if the local high school's percentage was lower. One hundred fifty students were chosen at random and surveyed. Of the 150 students surveyed, 82 have smoked. Use a significance level of 0.05 and using appropriate statistical evidence, conduct a hypothesis test and state the conclusion.

Exercise:

Problem:

A recent survey in the *N.Y. Times Almanac* indicated that 48.8% of families own stock. A broker wanted to determine if this survey could be valid. He surveyed a random sample of 250 families and found that 142 owned some type of stock. At the 0.05 significance level, can the survey be considered to be accurate?

Solution:

a. $H_0 : p = 0.488$ $H_a : p \neq 0.488$
b. p-value = 0.0114
c. alpha = 0.05
d. Reject the null hypothesis.
e. At the 5% level of significance, there is enough evidence to conclude that it is not true that 48.8% of families own stocks.
f. The survey does not appear to be accurate.

Exercise:

Problem:

Driver error can be listed as the cause of approximately 54% of all fatal auto accidents, according to the American Automobile Association. Thirty randomly selected fatal accidents are examined, and it is determined that 14 were caused by driver error. Using α = 0.05, is the AAA proportion accurate?

Exercise:

Problem:

The US Department of Energy reported that 51.7% of homes were heated by natural gas. A random sample of 221 homes in Kentucky found that 115 were heated by natural gas. Does the evidence support that this percentage holds true for Kentucky as well? Are the results applicable across the country? Why?

Solution:

a. $H_0 : p = 0.517$ $H_a : p \neq 0.517$
b. p-value = 0.9203.
c. alpha = 0.05.
d. Do not reject the null hypothesis.
e. At the 5% significance level, there is not enough evidence to conclude that the proportion of homes in Kentucky that are heated by natural gas is not 0.517.
f. However, we cannot generalize this result to the entire nation. First, the sample's population is only the state of Kentucky. Second, it is reasonable to assume that homes in the extreme north and south will have extreme high usage and low usage, respectively. We would need to expand our sample base to include these possibilities if we wanted to generalize this claim to the entire nation.

Exercise:

Problem:

For Americans using library services, the American Library Association claims that at most 67% of patrons borrow books. The library director in Owensboro, Kentucky feels this is not true, so she asked a local college statistic class to conduct a survey. The class randomly selected 100 patrons and found that 82 borrowed books. Did the class demonstrate that the percentage was higher in Owensboro, KY? Use $\alpha = 0.01$ level of significance. What is the possible proportion of patrons that do borrow books from the Owensboro Library?

Exercise:

Problem:

The Weather Underground reported that the mean amount of summer rainfall for the northeastern US is at least 11.52 inches. Ten cities in the northeast are randomly selected and the mean rainfall amount is calculated to be 7.42 inches with a standard deviation of 1.3 inches. At the $\alpha = 0.05$ level, can it be concluded that the mean rainfall was below the reported average? What if $\alpha = 0.01$? Assume the amount of summer rainfall follows a normal distribution.

Solution:

a. $H_0 : \mu = 11.52$ $H_a : \mu < 11.52$
b. p-value = 0.000002 which is almost 0.
c. alpha = 0.05.
d. Reject the null hypothesis.
e. At the 5% significance level, there is enough evidence to conclude that the mean amount of summer rain in the northeastern US is less than 11.52 inches, on average.
f. We would make the same conclusion if alpha was 1% because the p-value is almost 0.

Exercise:

Problem:

A survey in the *N.Y. Times Almanac* finds the mean commute time (one way) is 25.4 minutes for the 15 largest US cities. The Austin, TX chamber of commerce feels that Austin's commute time is less and wants to publicize this fact. The mean for 25 randomly selected commuters is 22.1 minutes with a standard deviation of 5.3 minutes. At the $\alpha = 0.05$ level, is the Austin, TX commute significantly less than the mean commute time for the 15 largest US cities?

Exercise:

Problem:

A report by the Gallup Poll found that a woman visits her doctor, on average, at most 5.8 times each year. A random sample of 20 women results in these yearly visit totals

3 2 1 3 7 2 9 4 6 6 8 0 5 6 4 2 1 3 4 1

At the $\alpha = 0.05$ level can it be concluded that the sample mean is higher than 5.8 visits per year?

Solution:

a. $H_0 : \mu = 5.8$ $H_a : \mu > 5.8$
b. p-value = 0.9987
c. alpha = 0.05
d. Do not reject the null hypothesis.
e. At the 5% level of significance, there is not enough evidence to conclude that a woman visits her doctor, on average, more than 5.8 times a year.

Exercise:

Problem:

According to the *N.Y. Times Almanac* the mean family size in the U.S. is 3.18. A sample of a college math class resulted in the following family sizes:
5 4 5 4 4 3 6 4 3 3 5 5 6 3 3 2 7 4 5 2 2 2 3 2

At $\alpha = 0.05$ level, is the class' mean family size greater than the national average? Does the Almanac result remain valid? Why?

Exercise:

Problem:

The student academic group on a college campus claims that freshman students study at least 2.5 hours per day, on average. One Introduction to Statistics class was skeptical and thinks this number is too high. The class took a random sample of 30 freshman students and found a mean study time of 137 minutes with a standard deviation of 45 minutes. At $\alpha = 0.01$ level, is the student academic group's claim correct?

Solution:

a. $H_0 : \mu = 150$ $H_a : \mu < 150$
b. p-value = 0.0622
c. alpha = 0.01
d. Do not reject the null hypothesis.
e. At the 1% significance level, there is not enough evidence to conclude that freshmen students study less than 2.5 hours per day, on average.
f. The student academic group's claim appears to be correct.

References

Data from Amit Schitai. Director of Instructional Technology and Distance Learning. LBCC.

Data from *Bloomberg Businessweek*. Available online at http://www.businessweek.com/news/2011- 09-15/nyc-smoking-rate-falls-to-record-low-of-14-bloomberg-says.html.

Data from energy.gov. Available online at http://energy.gov (accessed June 27. 2013).

Data from Gallup®. Available online at www.gallup.com (accessed June 27, 2013).

Data from *Growing by Degrees* by Allen and Seaman.

Data from La Leche League International. Available online at http://www.lalecheleague.org/Law/BAFeb01.html.

Data from the American Automobile Association. Available online at www.aaa.com (accessed June 27, 2013).

Data from the American Library Association. Available online at www.ala.org (accessed June 27, 2013).

Data from the Bureau of Labor Statistics. Available online at http://www.bls.gov/oes/current/oes291111.htm.

Data from the Centers for Disease Control and Prevention. Available online at www.cdc.gov (accessed June 27, 2013)

Data from the U.S. Census Bureau, available online at http://quickfacts.census.gov/qfd/states/00000.html (accessed June 27, 2013).

Data from the United States Census Bureau. Available online at http://www.census.gov/hhes/socdemo/language/.

Data from Toastmasters International. Available online at http://toastmasters.org/artisan/detail.asp?CategoryID=1&SubCategoryID=10&ArticleID=429&Page=1.

Data from Weather Underground. Available online at www.wunderground.com (accessed June 27, 2013).

Federal Bureau of Investigations. “Uniform Crime Reports and Index of Crime in Daviess in the State of Kentucky enforced by Daviess County from 1985 to 2005.” Available online at http://www.disastercenter.com/kentucky/crime/3868.htm (accessed June 27, 2013).

“Foothill-De Anza Community College District.” De Anza College, Winter 2006. Available online at http://research.fhda.edu/factbook/DAdemofs/Fact_sheet_da_2006w.pdf.

Johansen, C., J. Boice, Jr., J. McLaughlin, J. Olsen. “Cellular Telephones and Cancer—a Nationwide Cohort Study in Denmark.” Institute of Cancer Epidemiology and the Danish Cancer Society, 93(3):203-7. Available online at http://www.ncbi.nlm.nih.gov/pubmed/11158188 (accessed June 27, 2013).

Rape, Abuse & Incest National Network. “How often does sexual assault occur?” RAINN, 2009. Available online at http://www.rainn.org/get-information/statistics/frequency-of-sexual-assault (accessed June 27, 2013).

Glossary

Central Limit Theorem

> Given a random variable (RV) with known mean μ and known standard deviation σ. We are sampling with size n and we are interested the sample mean, X. If the size n of the sample is sufficiently large, then $X \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$. If the size n of the sample is sufficiently large, then the distribution of the sample means will approximate a normal distribution regardless of the shape of the population. The mean of the sample means will equal the population mean. The standard deviation of the distribution of the sample means, $\frac{\sigma}{\sqrt{n}}$, is called the standard error of the mean.

Hypothesis Testing With Two Samples: Introduction class="introduction"

If you want to test a claim that involves two groups (the types of breakfasts eaten east and west of the Mississippi River) you can use a slightly different technique when conducting a hypothesis test. (credit: Chloe Lim)

Note:
Chapter Objectives
By the end of this chapter, the student should be able to:

- Classify hypothesis tests by type.
- Conduct and interpret hypothesis tests for two population means, population standard deviations unknown.
- Conduct and interpret hypothesis tests for two population proportions.
- Conduct and interpret hypothesis tests for matched or paired samples.

Studies often compare two groups. For example, researchers are interested in the effect aspirin has in preventing heart attacks. Over the last few years, newspapers and magazines have reported various aspirin studies involving two groups. Typically, one group is given aspirin and the other group is given a placebo. Then, the heart attack rate is studied over several years.

There are other situations that deal with the comparison of two groups. For example, studies compare various diet and exercise programs. Politicians compare the proportion of individuals from different income brackets who might vote for them. Students are interested in whether SAT or GRE preparatory courses really help raise their scores.

You have learned to conduct hypothesis tests on single means and single proportions. You will expand upon that in this chapter. You will compare two means or two proportions to each other. The general procedure is still the same, just expanded.

To compare two means or two proportions, you work with two groups. The groups are classified either as **independent** or **matched pairs**. **Independent groups** consist of two samples that are independent, that is, sample values selected from one population are not related in any way to sample values selected from the other population. **Matched pairs** consist of two samples that are dependent, that is each individual in one sample is paired up with an individual in the second sample in some meaningful way. The parameter tested using matched pairs is the population mean of differences. The parameters tested using independent groups are either population means or population proportions.

Note:

NOTE

This chapter relies on either a calculator or a computer to calculate the degrees of freedom, the test statistics, and *p*-values. TI-83+ and TI-84 instructions are included as well as the test statistic formulas. When using a TI-83+ or TI-84 calculator, we do not need to separate two population means, independent groups, or unknown population variances into large and small sample sizes. However, most statistical computer software has the ability to differentiate these tests.

This chapter deals with the following hypothesis tests:

Independent groups (samples are independent)

- Test of two population means.
- Test of two population proportions.

Matched or paired samples (samples are dependent)

- Test of two population means by testing one population mean of differences.

Two Population Means with Unknown Standard Deviations

Sometimes we want to compare two population means. We may be looking for any difference between the means or we may be looking to see if one mean is more or less than the other.

If we run a hypothesis test to compare two population means by collecting data from two **independent samples**, then the hypotheses will have one of the following forms:

$H_0 : \mu_1 = \mu_2$ or equivalently $H_0 : \mu_1 - \mu_2 = 0$
$H_a : \mu_1 \neq \mu_2$ or equivalently $H_a : \mu_1 - \mu_2 \neq 0$

$H_0 : \mu_1 = \mu_2$ or equivalently $H_0 : \mu_1 - \mu_2 = 0$
$H_a : \mu_1 > \mu_2$ or equivalently $H_a : \mu_1 - \mu_2 > 0$

$H_0 : \mu_1 = \mu_2$ or equivalently $H_0 : \mu_1 - \mu_2 = 0$
$H_a : \mu_1 < \mu_2$ or equivalently $H_a : \mu_1 - \mu_2 < 0$

Since population standard deviations are rarely known, we will limit our tests to compare two population means to those with unknown population standard deviations.

1. The two **independent samples** are simple random samples from two distinct populations.
2. For the two distinct populations:
    - if the sample sizes are small, the distributions are important (should be normal)
    - if the sample sizes are large, the distributions are not important (need not be normal)

Recall that two samples are independent when sample values selected from one population are not related in any way to sample values selected from the other population.

Note: The test comparing two independent population means with unknown and possibly unequal population standard deviations is called the Aspin-Welch t-test. The degrees of freedom formula was developed by Aspin-Welch.

The comparison of two population means is very common. A difference between the two samples depends on both the means and the standard deviations. Very different means can occur by chance if there is great variation among the individual samples. In order to account for the variation, we take the difference of the sample means, $\overline{X}_1 - \overline{X}_2$, and divide by the standard error in order to standardize the difference. The result is a t-score test statistic.

Because we do not know the population standard deviations, we estimate them using the two sample standard deviations from our independent samples. For the hypothesis test, we calculate

the estimated standard deviation, or **standard error**, of **the difference in sample means**, $\overline{X}_1$ – $\overline{X}_2$.
Equation:

The standard error is:

$$\sqrt{\frac{(s_1)^2}{n_1}+\frac{(s_2)^2}{n_2}}$$

The test statistic (t-score) is calculated as follows:
Equation:

$$\frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{(s_1)^2}{n_1}+\frac{(s_2)^2}{n_2}}}$$

where:

- s_1 and s_2, the sample standard deviations, are estimates of σ_1 and σ_2, respectively.
- σ_1 and σ_2 are the unknown population standard deviations.
- $\overline{x}_1$ and $\overline{x}_2$ are the sample means. μ_1 and μ_2 are the population means.

The number of **degrees of freedom (df)** requires a somewhat complicated calculation. However, a computer or calculator calculates it easily. The df are not always a whole number. The test statistic calculated previously is approximated by the Student's t-distribution with df as follows:
Equation:

Degrees of freedom

$$df = \frac{\left(\frac{(s_1)^2}{n_1}+\frac{(s_2)^2}{n_2}\right)^2}{\left(\frac{1}{n_1-1}\right)\left(\frac{(s_1)^2}{n_1}\right)^2+\left(\frac{1}{n_2-1}\right)\left(\frac{(s_2)^2}{n_2}\right)^2}$$

When both sample sizes n_1 and n_2 are five or larger, the Student's t approximation is very good. Notice that the sample variances $(s_1)^2$ and $(s_2)^2$ are not pooled. (If the question comes up, do not pool the variances.)

Note: It is not necessary to compute this by hand. A calculator or computer easily computes it.

Example:
Independent groups
The average amount of time boys and girls aged seven to 11 spend playing sports each day is believed to be the same. A study is done and data are collected, resulting in the data in the

following table. Each population has a normal distribution.

	Sample Size	Average Number of Hours Playing Sports Per Day	Sample Standard Deviation
Girls	9	2	0.866
Boys	16	3.2	1.00

Exercise:

Problem:

Is there a difference in the mean amount of time boys and girls aged seven to 11 play sports each day? Test at the 5% level of significance.

Solution:

The population standard deviations are not known. Let g be the subscript for girls and b be the subscript for boys. Then, μ_g is the population mean for girls and μ_b is the population mean for boys. This is a test of two **independent groups**, two population **means**.

Random variable: $\overline{X}_g - \overline{X}_b$ = difference in the sample mean amount of time girls and boys play sports each day.

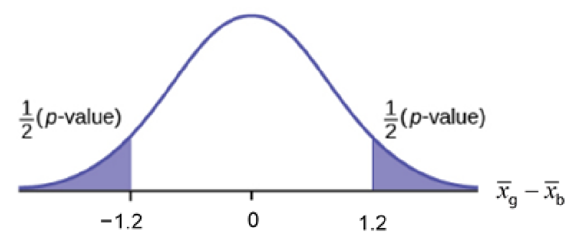
$H_0 : \mu_g = \mu_b$ $\qquad H_0 : \mu_g - \mu_b = 0$
$H_a : \mu_g \neq \mu_b$ $\qquad H_a : \mu_g - \mu_b \neq 0$

Since there are no words to indicate if you are testing for evidence that one population mean is greater (or less) than the other, assume you are looking for any difference and use the not equal sign for H_a . This is a two-tailed test.

Distribution for the test: Use t_{df} where df is calculated using the df formula for independent groups, two population means. Using a calculator, df is approximately 18.8466. **Do not pool the variances.**

Calculate the p-value using a Student's t-distribution: p-value = 0.0054

Graph:



$s_g = 0.866$

$s_b = 1$

So, $\overline{x}_g - \overline{x}_b$ = 2 – 3.2 = –1.2

Half the p-value is below –1.2 and half is above 1.2.

Make a decision: Since $p < \alpha$, reject H_0. This means you reject $\mu_g = \mu_b$. The means are different.

Note:
Press `STAT`. Arrow over to `TESTS` and press `4:2-SampTTest`. Arrow over to Stats and press `ENTER`. Arrow down and enter `2` for the first sample mean, `0.866` for Sx1, `9` for n1, `3.2` for the second sample mean, `1` for Sx2, and `16` for n2. Arrow down to μ1: and arrow to `does not equal` μ2. Press `ENTER`. Arrow down to Pooled: and `No`. Press `ENTER`. Arrow down to `Calculate` and press `ENTER`. The p-value is p = 0.0054, the df are approximately 18.8466, and the test statistic is -3.14. Do the procedure again but instead of Calculate do Draw.

Conclusion: At the 5% level of significance, the sample data show there is sufficient evidence to conclude that the mean number of hours that girls and boys aged seven to 11 play sports per day is different.

Note:
Try It
Exercise:

Problem:

Two samples are shown in the following table. Both have normal distributions. The means for the two populations are thought to be the same. Is there a difference in the means? Test at the 5% level of significance.

	Sample Size	Sample Mean	Sample Standard Deviation
Population A	25	5	1
Population B	16	4.7	1.2

Solution:

The *p*-value is 0.4125, which is much higher than 0.05, so we decline to reject the null hypothesis. There is not sufficient evidence to conclude that the means of the two populations are different.

Example:
A study is done by a community group in two neighboring colleges to determine which one graduates students with more math classes. College A samples 11 graduates. Their average is four math classes with a standard deviation of 1.5 math classes. College B samples nine graduates. Their average is 3.5 math classes with a standard deviation of one math class. The community group believes that a student who graduates from college A **has taken more math classes,** on average than a student who graduates from college B. Both populations have a normal distribution. Test at a 1% significance level. Answer the following questions.

Exercise:

Problem: a. Is this a test of two means or two proportions?

Solution:

a. two means

Exercise:

Problem: b. Are the population standard deviations known or unknown?

Solution:

b. unknown

Exercise:

Problem: c. Which distribution do you use to perform the test?

Solution:

c. Student's t

Exercise:

Problem: d. What is the random variable?

Solution:

d. $\overline{X}_A - \overline{X}_B$

Exercise:

Problem:

e. What are the null and alternative hypotheses? Write the null and alternative hypotheses.

Solution:

e.

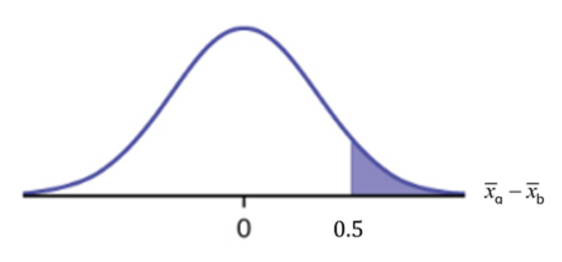
- $H_o : \mu_A = \mu_B$
- $H_a : \mu_A > \mu_B$

Exercise:

Problem: f. Is this test right-, left-, or two-tailed?

Solution:

f.



right

Exercise:

Problem:g. What is the *p*-value?

Solution:

g. 0.1928

Exercise:

Problem:h. Do you reject or not reject the null hypothesis?

Solution:

h. Do not reject.

Exercise:

Problem:i. **Conclusion:**

Solution:

i. At the 1% level of significance, from the sample data, there is not sufficient evidence to conclude that a student who graduates from college A has taken more math classes, on average, than a student who graduates from college B.

Note:
Try It
Exercise:

Problem:

A study is done to determine if Company A retains its workers longer than Company B. Company A randomly samples 15 workers, and their average time with the company is five years with a standard deviation of 1.2 years. Company B randomly samples 20 workers, and their average time with the company is 4.5 years with a standard deviation of 0.8 year. Assume the populations are normally distributed.

a. Are the population standard deviations known?
b. Conduct an appropriate hypothesis test. At the 5% significance level, what is your conclusion?

Solution:

a. They are unknown.
b. The p-value = 0.0878. At the 5% level of significance, there is insufficient evidence to conclude that the workers of Company A stay longer with the company than workers from Company B.

Example:

A professor at a large community college wanted to determine whether there is a difference in the means of final exam scores between students who took his statistics course online and the students who took his face-to-face statistics class. He believed that the mean of the final exam scores for the online class would be lower than that of the face-to-face class. Was the professor correct? The randomly selected 30 final exam scores from each group are listed in [link] and [link].

67.6	41.2	85.3	55.9	82.4	91.2	73.5	94.1	64.7	64.7
70.6	38.2	61.8	88.2	70.6	58.8	91.2	73.5	82.4	35.5
94.1	88.2	64.7	55.9	88.2	97.1	85.3	61.8	79.4	79.4

Online Class

77.9	95.3	81.2	74.1	98.8	88.2	85.9	92.9	87.1	88.2
69.4	57.6	69.4	67.1	97.6	85.9	88.2	91.8	78.8	71.8
98.8	61.2	92.9	90.6	97.6	100	95.3	83.5	92.9	89.4

Face-to-face Class

Exercise:

Problem:

Is the mean of the Final Exam scores of the online class lower than the mean of the Final Exam scores of the face-to-face class? Test at a 5% significance level. Answer the following questions:

a. Is this a test of two means or two proportions?
b. Are the population standard deviations known or unknown?
c. Which distribution do you use to perform the test?
d. What is the random variable?
e. What are the null and alternative hypotheses? Write the null and alternative hypotheses in words and in symbols.
f. Is this test right, left, or two tailed?
g. What is the p-value?
h. Do you reject or not reject the null hypothesis?
i. At the ___ level of significance, from the sample data, there _______ (is/is not) sufficient evidence to conclude that _______.

(See the conclusion in [link], and write yours in a similar fashion)

Note:
First put the data for each group into two lists (such as L1 and L2). Press STAT. Arrow over to TESTS and press 4:2SampTTest. Make sure Data is highlighted and press ENTER. Arrow down and enter L1 for the first list and L2 for the second list. Arrow down to μ_1: and arrow to $\neq \mu_2$ (does not equal). Press ENTER. Arrow down to Pooled: No. Press ENTER. Arrow down to Calculate and press ENTER.

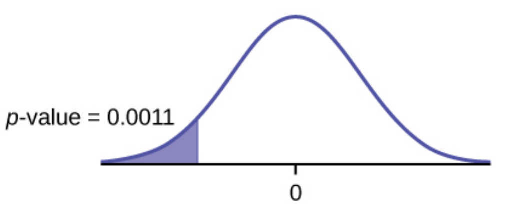
Note:
Note
Be careful not to mix up the information for Group 1 and Group 2!

Solution:

a. two means
b. unknown
c. Student's t
d. $\overline{X}_1 - \overline{X}_2$

e.
  - $H_0 : \mu_1 = \mu_2$ Null hypothesis: the means of the final exam scores are equal for the online and face-to-face statistics classes.
  - $H_a : \mu_1 < \mu_2$ Alternative hypothesis: the mean of the final exam scores of the online class is less than the mean of the final exam scores of the face-to-face class.

f. left-tailed
g. p-value = 0.0011



h. Reject the null hypothesis
i. The professor was correct. The evidence shows that the mean of the final exam scores for the online class is lower than that of the face-to-face class.
At the 5% level of significance, from the sample data, there is sufficient evidence to conclude that the mean of the final exam scores for the online class is less than the mean of final exam scores of the face-to-face class.

Note:
Try It
Weighted alpha is a measure of risk-adjusted performance of stocks over a period of a year. A high positive weighted alpha signifies a stock whose price has risen while a small positive weighted alpha indicates an unchanged stock price during the time period. Weighted alpha is used to identify companies with strong upward or downward trends. The weighted alpha for the top 30 stocks of banks in the northeast and in the west as identified by Nasdaq on May 24, 2013 are listed in [link] and [link], respectively.

94.2	75.2	69.6	52.0	48.0	41.9	36.4	33.4	31.5	27.6
77.3	71.9	67.5	50.6	46.2	38.4	35.2	33.0	28.7	26.5
76.3	71.7	56.3	48.7	43.2	37.6	33.7	31.8	28.5	26.0

Northeast

126.0	70.6	65.2	51.4	45.5	37.0	33.0	29.6	23.7	22.6
116.1	70.6	58.2	51.2	43.2	36.0	31.4	28.7	23.5	21.6
78.2	68.2	55.6	50.3	39.0	34.1	31.0	25.3	23.4	21.5

West

Exercise:

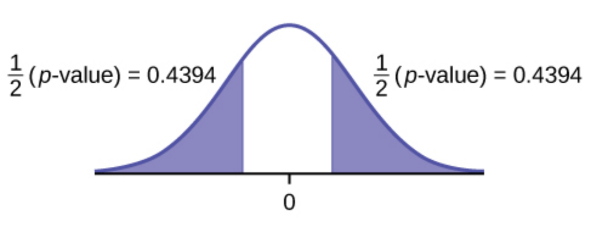
Problem:

Is there a difference in the weighted alpha of the top 30 stocks of banks in the northeast and in the west? Test at a 5% significance level. Answer the following questions:

a. Is this a test of two means or two proportions?
b. Are the population standard deviations known or unknown?
c. Which distribution do you use to perform the test?
d. What is the random variable?
e. What are the null and alternative hypotheses?
f. Is this test right, left, or two tailed?
g. What is the p-value?
h. Do you reject or not reject the null hypothesis?
i. At the ___ level of significance, from the sample data, there ______ (is/is not) sufficient evidence to conclude that ______.

Solution:

a. two means
b. unknown
c. Student’s t
d. $\overline{X}_1$ - $\overline{X}_2$
e.
   - $H_0 : \mu_1 = \mu_2$
   - $H_a : \mu_1 \neq \mu_2$
f. two-tailed
g. p-value = 0.8787
h. Do not reject the null hypothesis
i. This indicates that the trends in stocks are about the same in the top 30 banks in each region.



5% level of significance, from the sample data, there is not sufficient evidence to conclude that the mean weighted alphas for the banks in the northeast and the west are different

References

Data from Graduating Engineer + Computer Careers. Available online at http://www.graduatingengineer.com

Data from *Microsoft Bookshelf*.

Data from the United States Senate website, available online at www.Senate.gov (accessed June 17, 2013).

"List of current United States Senators by Age." Wikipedia. Available online at http://en.wikipedia.org/wiki/List_of_current_United_States_Senators_by_age (accessed June 17, 2013).

"Sectoring by Industry Groups." Nasdaq. Available online at http://www.nasdaq.com/markets/barchart-sectors.aspx?page=sectors&base=industry (accessed June 17, 2013).

"Strip Clubs: Where Prostitution and Trafficking Happen." Prostitution Research and Education, 2013. Available online at www.prostitutionresearch.com/ProsViolPosttrauStress.html (accessed June 17, 2013).

"World Series History." Baseball-Almanac, 2013. Available online at http://www.baseball-almanac.com/ws/wsmenu.shtml (accessed June 17, 2013).

Section Review

Two population means from independent samples where the population standard deviations are not known

- Random Variable: $\overline{X}_1 - \overline{X}_2$ = the difference of the sample means
- Distribution: Student's t-distribution with degrees of freedom (variances not pooled)

Formula Review

Standard error: $SE = \sqrt{\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}}$

Test statistic (t-score): $t = \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}}}$

Degrees of freedom:

$$df = \frac{\left(\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}\right)^2}{\left(\frac{1}{n_1 - 1}\right)\left(\frac{(s_1)^2}{n_1}\right)^2 + \left(\frac{1}{n_2 - 1}\right)\left(\frac{(s_2)^2}{n_2}\right)^2}$$

where s_1 and s_2 are the sample standard deviations, and n_1 and n_2 are the sample sizes.

You do not need to memorize the degrees of freedom formula. When you run a 2-SampTTest, your calculator will find the degrees of freedom for you.

$\overline{x}_1$ and $\overline{x}_2$ are the sample means.

Use the following information to answer the next 13 exercises: Indicate if the hypothesis test is for

a. independent group means, population standard deviations unknown
b. matched or paired samples
c. single mean
d. two proportions
e. single proportion

Exercise:

Problem:

It is believed that 70% of males pass their drivers test in the first attempt, while 65% of females pass the test in the first attempt. Of interest is whether the proportions are in fact equal.

Solution:

two proportions

Exercise:

Problem:

A new laundry detergent is tested on consumers. Of interest is the proportion of consumers who prefer the new brand over the leading competitor. A study is done to test this.

Exercise:

Problem:

A new windshield treatment claims to repel water more effectively. Ten windshields are tested by simulating rain without the new treatment. The same windshields are then treated, and the experiment is run again. A hypothesis test is conducted.

Solution:

matched or paired samples

Exercise:

Problem: The average worker in Germany gets eight weeks of paid vacation.

Solution:

single mean

Exercise:

Problem:

According to a television commercial, 80% of dentists agree that Ultrafresh toothpaste is the best on the market.

Exercise:

Problem:

It is believed that the average grade on an English essay in a particular school system for females is higher than for males. A random sample of 31 females had a mean score of 82 with a standard deviation of three, and a random sample of 25 males had a mean score of 76 with a standard deviation of four.

Solution:

independent group means, population standard deviations unknown

Exercise:

Problem:

In a random sample of 100 forests in the United States, 56 were coniferous or contained conifers. In a random sample of 80 forests in Mexico, 40 were coniferous or contained conifers. Is the proportion of conifers in the United States statistically more than the proportion of conifers in Mexico?

Solution:

two proportions

Exercise:

Problem:

A new medicine is said to help improve sleep. Eight subjects are picked at random and given the medicine. The means hours slept for each person were recorded before starting the medication and after.

Exercise:

Problem:

It is thought that teenagers sleep more than adults on average. A study is done to verify this. A sample of 16 teenagers has a mean of 8.9 hours slept and a standard deviation of 1.2. A sample of 12 adults has a mean of 6.9 hours slept and a standard deviation of 0.6.

Solution:

independent group means, population standard deviations unknown

Exercise:

Problem: Varsity athletes practice five times a week, on average.

Exercise:

Problem:

A sample of 12 in-state graduate school programs at school A has a mean tuition of $64,000 with a standard deviation of $8,000. At school B, a sample of 16 in-state graduate programs has a mean of $80,000 with a standard deviation of $6,000. On average, are the mean tuitions different?

Solution:

independent group means, population standard deviations unknown

Exercise:

Problem:

A new WiFi range booster is being offered to consumers. A researcher tests the native range of 12 different routers under the same conditions. The ranges are recorded. Then the researcher uses the new WiFi range booster and records the new ranges. Does the new WiFi range booster do a better job?

Exercise:

Problem:

A high school principal claims that 30% of student athletes drive themselves to school, while 4% of non-athletes drive themselves to school. In a sample of 20 student athletes, 45% drive themselves to school. In a sample of 35 non-athlete students, 6% drive themselves to school. Is the percent of student athletes who drive themselves to school more than the percent of nonathletes?

Solution:

two proportions

Use the following information to answer the next three exercises: A study is done to determine which of two soft drinks has more sugar. There are 13 cans of Beverage A in a sample and six cans of Beverage B. The mean amount of sugar in Beverage A is 36 grams with a standard deviation of 0.6 grams. The mean amount of sugar in Beverage B is 38 grams with a standard deviation of 0.8 grams. The researchers believe that Beverage B has more sugar than Beverage A, on average. Both populations have normal distributions.

Exercise:

Problem: Are standard deviations known or unknown?

Exercise:

Problem: What is the random variable?

Solution:

The random variable is the difference between the mean amounts of sugar in the two soft drinks.

Exercise:

Problem: Is this a one-tailed or two-tailed test?

Use the following information to answer the next 12 exercises: The U.S. Center for Disease Control reports that the mean life expectancy was 47.6 years for whites born in 1900 and 33.0 years for nonwhites. Suppose that you randomly survey death records for people born in 1900 in a certain county. Of the 124 whites, the mean life span was 45.3 years with a standard deviation of 12.7 years. Of the 82 nonwhites, the mean life span was 34.1 years with a standard deviation of 15.6 years. Conduct a hypothesis test to see if the mean life spans in the county were the same for whites and nonwhites.

Exercise:

Problem: Is this a test of means or proportions?

Solution:

means

Exercise:

Problem: State the null and alternative hypotheses.

a. H_0 : ___________
b. H_a : ___________

Exercise:

Problem: Is this a right-tailed, left-tailed, or two-tailed test?

Solution:

two-tailed

Exercise:

Problem: In symbols, what is the random variable of interest for this test?

Exercise:

Problem: In words, define the random variable of interest for this test.

Solution:

the difference between the mean life spans of whites and nonwhites

Exercise:

Problem: Which distribution (normal or Student's t) would you use for this hypothesis test?

Exercise:

Problem: Explain why you chose the distribution you did for [link].

Solution:

This is a comparison of two population means with unknown population standard deviations.

Exercise:

Problem: Calculate the test statistic and p-value.

Exercise:

Problem:

Sketch a graph of the situation. Label the horizontal axis. Mark the hypothesized difference and the sample difference. Shade the area corresponding to the p-value.

Solution:

Check student’s solution.

Exercise:

Problem: Find the *p*-value.

Exercise:

Problem: At a pre-conceived $\alpha = 0.05$, what is your:

a. Decision:
b. Reason for the decision:
c. Conclusion (write out in a complete sentence):

Solution:

a. Reject the null hypothesis
b. *p*-value < 0.05
c. There is evidence at the 5% level of significance to support the claim that life expectancy in the 1900s is different between whites and nonwhites.

Exercise:

Problem: Does it appear that the means are the same? Why or why not?

Homework

For each of the word problems below, use a solution sheet to do the hypothesis test. The Solutions Sheets can be found in the Table of Contents or by clicking here. Please feel free to make copies of the Solution Sheets. For the online version of the book, it is suggested that you copy the .doc or the .pdf files.

> **Note:**
> NOTE
> If you are using a Student's t-distribution for a homework problem in what follows, including for paired data, you may assume that the underlying population is normally distributed. (When using these tests in a real situation, you must first prove that assumption, however.)

Exercise:

Problem:

The mean number of English courses taken in a two–year time period by male and female college students is believed to be about the same. An experiment is conducted and data are collected from 29 males and 16 females. The males took an average of three English courses with a standard deviation of 0.8. The females took an average of four English courses with a standard deviation of 1.0. Are the means statistically the same?

Exercise:

Problem:

A student at a four-year college claims that mean enrollment at four–year colleges is higher than at two–year colleges in the United States. Two surveys are conducted. Of the 35 two–year colleges surveyed, the mean enrollment was 5,068 with a standard deviation of 4,777. Of the 35 four-year colleges surveyed, the mean enrollment was 5,466 with a standard deviation of 8,191.

Solution:

Subscripts: 1: two-year colleges; 2: four-year colleges

a. $H_0 : \mu_1 = \mu_2$
b. $H_a : \mu_1 < \mu_2$
c. $\overline{X}_1 - \overline{X}_2$ is the difference between the mean enrollments of the two-year colleges and the four-year colleges.
d. Student’s t
e. test statistic: t = -0.2480
f. p-value: 0.4019
g. Check student's solution.

h.
   i. Alpha: 0.05
   ii. Decision: Do not reject
   iii. Reason for Decision: p-value > alpha
   iv. Conclusion: At the 5% significance level, there is not sufficient evidence to conclude that the mean enrollment at four-year colleges is higher than at two-year colleges.

Exercise:

Problem:

At Rachel’s 11th birthday party, eight girls were timed to see how long (in seconds) they could hold their breath in a relaxed position. After a two-minute rest, they timed themselves while jumping. The girls thought that the mean difference between their jumping and relaxed times would be zero. Test their hypothesis.

Relaxed time (seconds)	Jumping time (seconds)
26	21
47	40
30	28
22	21
23	25
45	43
37	35
29	32

Exercise:

Problem:

Mean entry-level salaries for college graduates with mechanical engineering degrees and electrical engineering degrees are believed to be approximately the same. A recruiting office thinks that the mean mechanical engineering salary is actually lower than the mean electrical engineering salary. The recruiting office randomly surveys 50 entry level mechanical engineers and 60 entry level electrical engineers. Their mean salaries were $46,100 and $46,700, respectively. Their standard deviations were $3,450 and $4,210, respectively. Conduct a hypothesis test to determine if you agree that the mean entry-level mechanical engineering salary is lower than the mean entry-level electrical engineering salary.

Solution:

Subscripts: 1: mechanical engineering; 2: electrical engineering

a. $H_0 : \mu_1 = \mu_2$
b. $H_a : \mu_1 < \mu_2$
c. $\overline{X}_1 - \overline{X}_2$ is the difference between the mean entry level salaries of mechanical engineers and electrical engineers.
d. Student's t
e. test statistic: $t = -0.82$
f. p-value: 0.2061
g. Check student’s solution.
h.
    i. Alpha: 0.05
    ii. Decision: Do not reject the null hypothesis.
    iii. Reason for Decision: p-value > alpha

    iv. Conclusion: At the 5% significance level, there is insufficient evidence to conclude that the mean entry-level salaries of mechanical engineers is lower than that of electrical engineers.

Exercise:

Problem:

Marketing companies have collected data implying that teenage girls use more ring tones on their cellular phones than teenage boys do. In one particular study of 40 randomly chosen teenage girls and boys (20 of each) with cellular phones, the mean number of ring tones for the girls was 3.2 with a standard deviation of 1.5. The mean for the boys was 1.7 with a standard deviation of 0.8. Conduct a hypothesis test to determine if the means are approximately the same or if the girls' mean is higher than the boys' mean.

Use the information from Appendix C to answer the next four exercises.

Exercise:

Problem:

Using the data from Lap 1 only, conduct a hypothesis test to determine if the mean time for completing a lap in races is the same as it is in practices.

Solution:

a. $H_0 : \mu_1 = \mu_2$
b. $H_a : \mu_1 \neq \mu_2$
c. $\overline{X}_1 - \overline{X}_2$ is the difference between the mean times for completing a lap in races and in practices.
d. $t_{20.32}$
e. test statistic: $t = –4.70$
f. p-value: 0.0001
g. Check student's solution.
h.
    i. Alpha: 0.05
    ii. Decision: Reject the null hypothesis.
    iii. Reason for Decision: p-value < alpha
    iv. Conclusion: At the 5% significance level, there is sufficient evidence to conclude that the mean time for completing a lap in races is different from that in practices.

Exercise:

Problem: Repeat the test in the previous exercise, but use Lap 5 data this time.

Exercise:

Problem: Repeat the test in [link], but this time combine the data from Laps 1 and 5.

Solution:

a. $H_0 : \mu_1 = \mu_2$
b. $H_a : \mu_1 \neq \mu_2$
c. is the difference between the mean times for completing a lap in races and in practices.
d. $t_{40.94}$
e. test statistic: $t = –5.08$
f. p-value: zero
g. Check student’s solution.

h.
   i. Alpha: 0.05
   ii. Decision: Reject the null hypothesis.
   iii. Reason for Decision: p-value < alpha
   iv. Conclusion: At the 5% significance level, there is sufficient evidence to conclude that the mean time for completing a lap in races is different from that in practices.

Exercise:

Problem:

In two to three complete sentences, explain in detail how you might use Terri Vogel’s data to answer the following question. “Does Terri Vogel drive faster in races than she does in practices?”

Use the following information to answer the next two exercises. The Eastern and Western Major League Soccer conferences have a new Reserve Division that allows new players to develop their skills. Data for a randomly picked date showed the following annual goals.

Western	**Eastern**
Los Angeles 9	D.C. United 9
FC Dallas 3	Chicago 8
Chivas USA 4	Columbus 7
Real Salt Lake 3	New England 6
Colorado 4	MetroStars 5
San Jose 4	Kansas City 3

Conduct a hypothesis test to answer the next two exercises.

Exercise:

Problem: The **exact** distribution for the hypothesis test is:

a. the normal distribution
b. the Student's t-distribution
c. the uniform distribution
d. the exponential distribution

Exercise:

Problem: If the level of significance is 0.05, the conclusion is:

a. There is sufficient evidence to conclude that the **W** Division teams score fewer goals, on average, than the **E** teams
b. There is insufficient evidence to conclude that the **W** Division teams score more goals, on average, than the **E** teams.
c. There is insufficient evidence to conclude that the **W** teams score fewer goals, on average, than the **E** teams score.
d. Unable to determine

Solution:

c

Exercise:

Problem:

Suppose a statistics instructor believes that there is no significant difference between the mean class scores of statistics day students on Exam 2 and statistics night students on Exam 2. She takes random samples from each of the populations. The mean and standard deviation for 35 statistics day students were 75.86 and 16.91. The mean and standard deviation for 37 statistics night students were 75.41 and 19.73. The “day” subscript refers to the statistics day students. The “night” subscript refers to the statistics night students. A concluding statement is:

a. There is sufficient evidence to conclude that statistics night students' mean on Exam 2 is better than the statistics day students' mean on Exam 2.
b. There is insufficient evidence to conclude that the statistics day students' mean on Exam 2 is better than the statistics night students' mean on Exam 2.
c. There is insufficient evidence to conclude that there is a significant difference between the means of the statistics day students and night students on Exam 2.
d. There is sufficient evidence to conclude that there is a significant difference between the means of the statistics day students and night students on Exam 2.

Exercise:

Problem:

Researchers interviewed street prostitutes in Canada and the United States. The mean age of the 100 Canadian prostitutes upon entering prostitution was 18 with a standard deviation of six. The mean age of the 130 United States prostitutes upon entering prostitution was 20 with a standard deviation of eight. Is the mean age of entering prostitution in Canada lower than the mean age in the United States? Test at a 1% significance level.

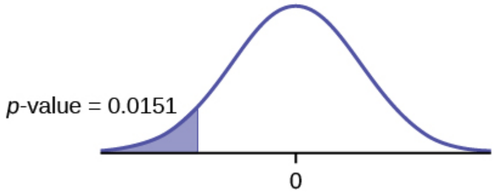
Solution:

Test: two independent sample means, population standard deviations unknown.

Random variable:

$$\overline{X}_1 - \overline{X}_2$$

Hypotheses: $H_0 : \mu_1 = \mu_2$; $H_a : \mu_1 < \mu_2$; The mean age of entering prostitution in Canada is lower than the mean age in the United States.



Graph: left-tailed

p-value : 0.0151

Decision: Do not reject H_0.

Conclusion: At the 1% level of significance, from the sample data, there is not sufficient evidence to conclude that the mean age of entering prostitution in Canada is lower than the mean age in the United States.

Exercise:

Problem:

A powder diet is tested on 49 people, and a liquid diet is tested on 36 different people. Of interest is whether the liquid diet yields a higher mean weight loss than the powder diet. The powder diet group had a mean weight loss of 42 pounds with a standard deviation of 12 pounds. The liquid diet group had a mean weight loss of 45 pounds with a standard deviation of 14 pounds.

Exercise:

Problem:

Suppose a statistics instructor believes that there is no significant difference between the mean class scores of statistics day students on Exam 2 and statistics night students on Exam 2. She takes random samples from each of the populations. The mean and standard deviation for 35 statistics day students were 75.86 and 16.91, respectively. The mean and standard deviation for 37 statistics night students were 75.41 and 19.73. The “day” subscript refers to the statistics day students. The “night” subscript refers to the statistics night students. An appropriate alternative hypothesis for the hypothesis test is:

a. $\mu_{\text{day}} > \mu_{\text{night}}$
b. $\mu_{\text{day}} < \mu_{\text{night}}$
c. $\mu_{\text{day}} = \mu_{\text{night}}$
d. $\mu_{\text{day}} \neq \mu_{\text{night}}$

Solution:

d

Glossary

Independent samples
: samples that are independent, that is, the sample values selected from one population are not related in any way to the sample values selected from the other population.

Comparing Two Independent Population Proportions

When conducting a hypothesis test that compares two independent population proportions, the following characteristics should be present:

1. The two independent samples are simple random samples that are independent.
2. The number of successes is at least ten, and the number of failures is at least ten, for each of the samples.
3. Growing literature states that the population must be at least ten or 20 times the size of the sample. This keeps each population from being over-sampled and causing incorrect results.

Comparing two proportions, like comparing two means, is common. If two estimated proportions are different, it may be due to a difference in the populations or it may be due to chance. A hypothesis test can help determine if a difference in the estimated proportions reflects a difference in the population proportions.

The difference of two proportions follows an approximate normal distribution. Generally, the null hypothesis states that the two proportions are the same. That is, $H_0 : p_{\mathrm{A}} = p_{\mathrm{B}}$, or equivalently, $H_0 : p_{\mathrm{A}} - p_{\mathrm{B}} = 0$. To conduct the test, we use a pooled proportion, p_{c} , which is an estimate of the common value the two population proportions are equal to, assuming the null hypothesis is true.

Equation:

The pooled proportion is calculated as follows:

$$p_c = \frac{x_A + x_B}{n_A + n_B}$$

where x_A and x_B are the number of successes in groups A and B, respectively, and n_A and n_B are the samples sizes of groups A and B, respectively.

The distribution is

$$\widehat{P}_A - \widehat{P}_B \sim N\left(0, \sqrt{p_c\left(1 - p_c\right)\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}\right)$$

Equation:

The test statistic is

$$z = \frac{\widehat{p}_A - \widehat{p}_B}{\sqrt{p_c\left(1 - p_c\right)\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}}$$

Example:
Exercise:

Problem:

Two types of medication for hives are being tested to determine if there is a **difference in the proportions of adult patient reactions. Twenty** out of a random **sample of 200** adults given medication A still had hives 30 minutes after taking the medication. **Twelve** out of another **random sample of 200 adults** given medication B still had hives 30 minutes after taking the medication. Test at a 1% level of significance.

Solution:

The problem asks for a difference in proportions, making it a test of two proportions.

Let A and B be the subscripts for medication A and medication B, respectively. Then p_{A} and p_{B} are the desired population proportions.

Random Variable:
$\widehat{P}_{\text{A}} - \widehat{P}_{\text{B}}$ = difference in the proportions of adult patients who did not react after 30 minutes to medication A and to medication B.

$H_0 : p_{\text{A}} = p_{\text{B}}$, or equivalently, $H_0 : p_{\text{A}} - p_{\text{B}} = 0$

$H_a : p_{\text{A}} \neq p_{\text{B}}$, or equivalently, $H_a : p_{\text{A}} - p_{\text{B}} \neq 0$

The words **"is a difference"** tell you the test is two-tailed.

Distribution for the test: Since this is a test of two population proportions, the distribution is normal:

$$p_c = \frac{x_A+x_B}{n_A+n_B} = \frac{20+12}{200+200} = 0.08 \quad 1-p_c = 0.92$$

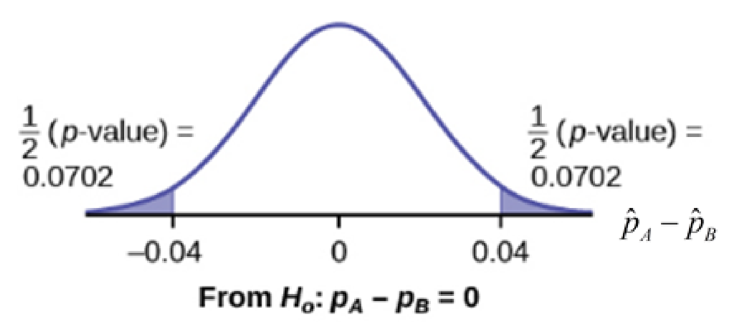
$$\widehat{p}_A - \widehat{p}_B \ \sim N\left(0, \sqrt{(0.08)(0.92)\left(\frac{1}{200} + \frac{1}{200}\right)}\right)$$

$\widehat{p}_A - \widehat{p}_B$ follows an approximate normal distribution.

Calculate the p-value using the normal distribution:
p-value = 0.1404.

Estimated proportion for group A: $\widehat{p}_A = \frac{x_A}{n_A} = \frac{20}{200} = 0.1$

Estimated proportion for group B: $\widehat{p}_B = \frac{x_B}{n_B} = \frac{12}{200} = 0.06$



$\hat{p}_A - \hat{p}_B = 0.1 - 0.06 = 0.04$.

Half the *p*-value is below –0.04, and half is above 0.04.

Compare α and the *p*-value: $\alpha = 0.01$ and the *p*-value = 0.1404.
So the *p*-value > α.

Make a decision: Since *p*-value > α, do not reject H_0 .

Conclusion: At a 1% level of significance, from the sample data, there is not sufficient evidence to conclude that there is a difference in the proportions of adult patients who did not react after 30 minutes to medication *A* and medication *B*.

Note:
Press `STAT`. Arrow over to `TESTS` and press `6:2-PropZTest`. Arrow down and enter `20` for x1, `200` for n1, `12` for x2, and `200` for n2. Arrow down to `p1`: and arrow to `≠ p2`. Press `ENTER`. Arrow down to `Calculate` and press `ENTER`. The *p*-value is $p = 0.1404$ and the test statistic is 1.47. Do the procedure again, but instead of `Calculate` do `Draw`.

Note:
Try It
Exercise:

Problem:

Two types of valves are being tested to determine if there is a difference in pressure tolerances. Twenty-two out of a random sample of 100 of Valve *A* cracked under 4,500 psi. Eleven out of a random sample of 100 of Valve *B* cracked under 4,500 psi. Test at a 5% level of significance.

Solution:

The *p*-value is 0.0361, so we can reject the null hypothesis. At the 5% significance level, the data support that there is a difference in the pressure tolerances between the two valves.

Example:
Exercise:

Problem:

A research study was conducted about gender differences in “sexting.” The researcher believed that the proportion of girls involved in “sexting” is less than the proportion of boys involved. The data collected in the spring of 2010 among a random sample of middle and high school students in a large school district in the southern United States is summarized in the table below. Is the proportion of girls sending sexts less than the proportion of boys “sexting?” Test at a 1% level of significance.

	Males	Females
Sent “sexts”	183	156
Total number surveyed	2231	2169

Solution:

This is a test of two population proportions. Let M and F be the subscripts for males and females. Then p_{M} and p_{F} are the desired population proportions.

Random variable:
$\widehat{P}_{\text{F}} - \widehat{P}_{\text{M}}$ = difference in the proportions of males and females who sent “sexts.”

$H_0 : p_{\text{F}} = p_{\text{M}}$, or equivalently, $H_0 : p_{\text{F}} - p_{\text{M}} = 0$

$H_a : p_{\text{F}} < p_{\text{M}}$, or equivalently, $H_a : p_{\text{F}} - p_{\text{M}} < 0$

The words **"less than"** tell you the test is left-tailed.

Distribution for the test: Since this is a test of two population proportions, the distribution is normal:

$$p_c = \frac{x_F + x_M}{n_F + n_M} = \frac{156+183}{2169+2231} = 0.077$$

$$1 - p_c = 0.923$$

Therefore,

$$\widehat{p}_F - \widehat{p}_M \sim N\left(0, \sqrt{(0.077)(0.923)\left(\frac{1}{2169} + \frac{1}{2231}\right)}\right)$$

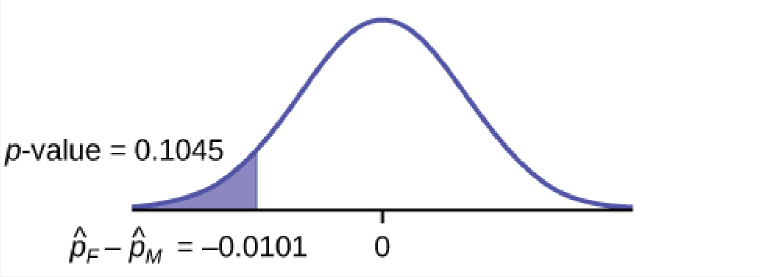
$\widehat{p}_F - \widehat{p}_M$ follows an approximate normal distribution.

Calculate the p-value using the normal distribution:

p-value = 0.1045

Estimated proportion for females: 0.0719

Estimated proportion for males: 0.082



Decision: Since p-value > α, do not reject H_0

Conclusion: At the 1% level of significance, from the sample data, there is not sufficient evidence to conclude that the proportion of girls sending “sexts” is less than the proportion of boys sending “sexts.”

Note:
Press `STAT`. Arrow over to `TESTS` and press `6:2-PropZTest`. Arrow down and enter `156` for x1, `2169` for n1, `183` for x2, and `2231` for n2. Arrow down to `p1:` and arrow to `< p2`. Press `ENTER`. Arrow down to `Calculate` and press `ENTER`. The p-value is p = 0.1045 and the test statistic is z = -1.256.

Example:
Exercise:

Problem:

Researchers conducted a study of smartphone use among adults. A cell phone company claimed that iPhone smartphones are more popular with whites (non-Hispanic) than with African Americans. The results of the survey indicate that of the 232 African American cell phone owners randomly sampled, 5% have an iPhone. Of the 1,343 white cell phone owners randomly sampled, 10% own an iPhone. Test at the 5% level of significance. Is the proportion of white iPhone owners greater than the proportion of African American iPhone owners?

Solution:

This is a test of two population proportions. Let W and A be the subscripts for the whites and African Americans. Then p_{W} and p_{A} are the desired population proportions.

Random variable:
$\widehat{P}_{\mathrm{W}} - \widehat{P}_{\mathrm{A}}$ = difference in the proportions of whites and African Americans iPhone users.

$H_0 : p_{\mathrm{W}} = p_{\mathrm{A}}$, or equivalently, $H_0 : p_{\mathrm{W}} - p_{\mathrm{A}} = 0$

$H_a : p_{\mathrm{W}} > p_{\mathrm{A}}$, or equivalently, $H_a : p_{\mathrm{W}} - p_{\mathrm{A}} > 0$

The words "more popular" indicate that the test is right-tailed.

Distribution for the test: The distribution is approximately normal:

$$p_c = \frac{x_W + x_A}{n_W + n_A} = \frac{134+12}{1343+232} = 0.0927$$

$$1 - p_c = 0.9073$$

Therefore,

$\widehat{p}_W - \widehat{p}_A$ follows an approximate normal distribution as follows.

$$\widehat{p}_W - \widehat{p}_A \sim N\left(0, \sqrt{(0.0927)(0.9073)\left(\frac{1}{1343} + \frac{1}{232}\right)}\right)$$

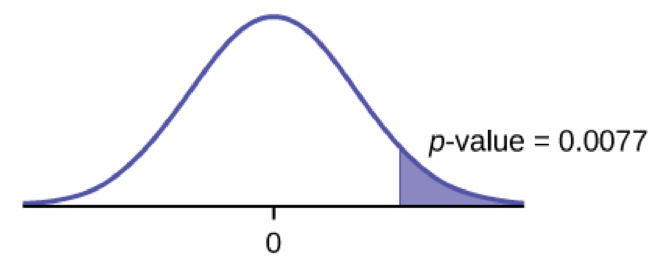
Test statistic
Equation:

$$z = \frac{0.10 - 0.05}{\sqrt{(0.0927)(0.9073)\left(\frac{1}{1343} + \frac{1}{232}\right)}} \approx 2.425$$

Calculate the p-value using the normal distribution:
p-value = 0.0077

Estimated proportion for group W: 0.10

Estimated proportion for group A: 0.05



Decision: Since p-value < α-value, reject the H_0 .

Conclusion: At the 5% level of significance, from the sample data, there is sufficient evidence to conclude that a larger proportion of white cell phone owners use iPhones than African Americans.

Note:
TI-83+ and TI-84: Press `STAT`. Arrow over to `TESTS` and press `6:2-PropZTest`. Arrow down and enter `134` for x1, `1343` for n1, `12` for x2, and `232` for n2. Arrow down to `p1:` and arrow to `> p2`. Press `ENTER`. Arrow down to `Calculate` and press `ENTER`. The p -value is $p = 0.0099$ and the test statistic is $z = 2.33$. (Note: The difference between this answer and the answer obtained previously is due to rounding.)

Note:
Try It
A concerned group of citizens wanted to know if the proportion of forcible rapes in Texas was different in 2011 than in 2010. Their research showed that of the 113,231 violent crimes in Texas in 2010, 7,622 of them were forcible rapes. In 2011, 7,439 of the 104,873 violent crimes were in the

forcible rape category. Test at a 5% significance level. Answer the following questions:

Exercise:

Problem:a. Is this a test of two means or two proportions?

Solution:

a. two proportions

Exercise:

Problem:b. Which distribution do you use to perform the test?

Solution:

b. normal for two proportions

Exercise:

Problem:c. What is the random variable?

Solution:

c. Subscripts: 1 = 2010, 2 = 2011

$\widehat{P}_1 - \widehat{P}_2$ = the difference in the proportion of forcible rapes in Texas in 2010 and in 2011.

Exercise:

Problem:

d. What are the null and alternative hypothesis? Write the null and alternative hypothesis in symbols.

Solution:

d. Subscripts: 1 = 2010, 2 = 2011

$H_0 : p_1 = p_2$, or equivalently, $H_0 : p_1 - p_2 = 0$
$H_a : p_1 \neq p_2$, or equivalently, $H_a : p_1 - p_2 \neq 0$

Exercise:

Problem:e. Is this test right-, left-, or two-tailed?

Solution:

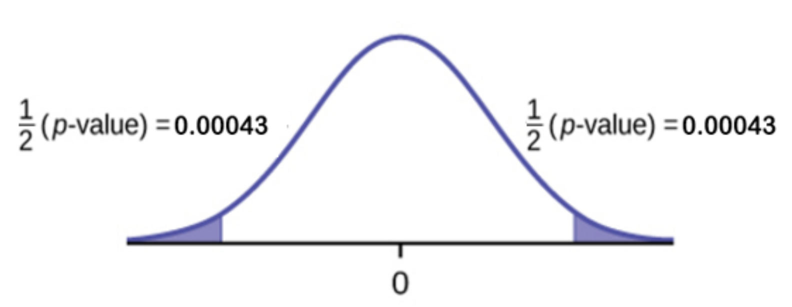
e. two-tailed

Exercise:

Problem:f. What is the *p*-value?

Solution:

f. *p*-value = 0.00086 (using the 2-PropZTest on the calculator)



Exercise:

Problem:g. Do you reject or not reject the null hypothesis?

Solution:

g. Reject H_0 .

Exercise:

Problem:

h. At the ___ level of significance, from the sample data, there ______ (is/is not) sufficient evidence to conclude that ____________.

Solution:

h. At the 5% level of significance, from the sample data, there is sufficient evidence to conclude that there is a difference between the proportion of forcible rapes in 2010 and 2011.

References

Data from *Educational Resources*, December catalog.

Data from Hilton Hotels. Available online at http://www.hilton.com (accessed June 17, 2013).

Data from Hyatt Hotels. Available online at http://hyatt.com (accessed June 17, 2013).

Data from Statistics, United States Department of Health and Human Services.

Data from Whitney Exhibit on loan to San Jose Museum of Art.

Data from the American Cancer Society. Available online at http://www.cancer.org/index (accessed June 17, 2013).

Data from the Chancellor’s Office, California Community Colleges, November 1994.

“State of the States.” Gallup, 2013. Available online at http://www.gallup.com/poll/125066/State-States.aspx?ref=interactive (accessed June 17, 2013).

“West Nile Virus.” Centers for Disease Control and Prevention. Available online at http://www.cdc.gov/ncidod/dvbid/westnile/index.htm (accessed June 17, 2013).

Section Review

Test of two population proportions from independent samples.

- Random variable: $\hat{P}_A - \hat{P}_B =$ difference between the two estimated proportions
- Distribution: normal distribution for two proportions

Formula Review

Pooled Proportion: $p_c = \frac{x_A + x_B}{n_A + n_B}$

Distribution for the differences:

$$\hat{P}_A - \hat{P}_B \sim N\left(0, \sqrt{p_c\left(1 - p_c\right)\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}\right)$$

Test Statistic $z = \frac{\hat{p}_A - \hat{p}_B}{\sqrt{p_c(1-p_c)\left(\frac{1}{n_A}+\frac{1}{n_B}\right)}}$

where the null hypothesis is $H_0 : p_A = p_B$, or equivalently, $H_0 : p_A - p_B = 0$.

$\hat{p}_A$ and $\hat{p}_B$ are the sample proportions, p_A and p_B are the population proportions,

p_c is the pooled proportion, and n_A and n_B are the sample sizes.

Use the following information for the next five exercises. Two types of phone operating system are being tested to determine if there is a difference in the proportions of system failures (crashes). Fifteen out of a random sample of 150 phones with OS_1 had system failures within the first eight hours of operation. Nine out of another random sample of 150 phones with OS_2 had system failures within the first eight hours of operation. OS_2 is believed to be more stable (have fewer crashes) than OS_1. Is there evidence to support this belief?

Exercise:

Problem: Is this a test of means or proportions?

Exercise:

Problem: What is the random variable?

Solution:

$\widehat{p}_{\text{OS1}} - \widehat{p}_{\text{OS2}}$ = difference in the proportions of phones that had system failures within the first eight hours of operation with OS_1 and OS_2.

Exercise:

Problem: State the null and alternative hypotheses.

Exercise:

Problem: What is the p-value?

Solution:

0.1018

Exercise:

Problem: What can you conclude about the two operating systems?

Use the following information to answer the next twelve exercises. In the recent Census, three percent of the U.S. population reported being of two or more races. However, the percent varies tremendously from state to state. Suppose that two random surveys are conducted. In the first random survey, out of 1,000 North Dakotans, only nineteen people reported being of two or more races. In the second random survey, out of 500 Nevadans, 17 people reported being of two or more races. Conduct a hypothesis test to determine if the population percents are the same for the two states or if the percent for Nevada is statistically higher than for North Dakota.

Exercise:

Problem: Is this a test of means or proportions?

Solution:

proportions

Exercise:

Problem: State the null and alternative hypotheses.

a. H_0 : __________
b. H_a : __________

Exercise:

Problem:

Is this a right-tailed, left-tailed, or two-tailed test? How do you know?

Solution:

right-tailed

Exercise:

Problem: What is the random variable of interest for this test?

Exercise:

Problem: In words, define the random variable for this test.

Solution:

The random variable is the difference in proportions (percents) of samples taken from the populations that are of two or more races in Nevada and North Dakota.

Exercise:

Problem: Which distribution would you use for this hypothesis test?

Exercise:

Problem:

Explain why you chose the distribution you did for the previous exercise.

Solution:

We are comparing two population proportions using independent samples. Also, the number of successes and failures in both samples are greater than ten each, so we can use the normal for two proportions distribution for this hypothesis test.

Exercise:

Problem: Calculate the test statistic.

Exercise:

Problem:

Sketch a graph of the situation. Mark the hypothesized difference and the sample difference. Shade the area corresponding to the p-value.

Solution:

Check student’s solution.

Exercise:

Problem: Find the p-value.

Exercise:

Problem: At a pre-conceived $\alpha = 0.05$, what is your:

a. Decision:
b. Reason for the decision:
c. Conclusion (write out in a complete sentence):

Solution:

a. Reject the null hypothesis.
b. The p-value < α.
c. At the 5% significance level, there is sufficient evidence to conclude that the proportion (percent) of the population that is of two or more races in Nevada is statistically higher than that in North Dakota.

Exercise:

Problem:

Does it appear that the proportion of Nevadans who are two or more races is higher than the proportion of North Dakotans? Why or why not?

Homework

For each of the word problems below, use a solution sheet to do the hypothesis test. The Solutions Sheets can be found in the Table of Contents or by clicking here. Please feel free to make copies of the Solution Sheets. For the online version of the book, it is suggested that you copy the .doc or the .pdf files.

Exercise:

Problem:

A recent drug survey showed an increase in the use of drugs and alcohol among local high school seniors as compared to the national percent. Suppose that a survey of 100 local seniors and 100 national seniors is conducted to see if the proportion of drug and alcohol use is higher locally than nationally. Locally, 65 seniors reported using drugs or alcohol within the past month, while 60 national seniors reported using them.

Exercise:

Problem:

We are interested in whether the proportions of female suicide victims for ages 15 to 24 are the same for the whites and the blacks races in the United States. We randomly pick one year, 1992, to compare the races. The number of suicides estimated in the United States in 1992 for white females is 4,930. Five hundred eighty were aged 15 to 24. The estimate for black females is 330. Forty were aged 15 to 24. We will let female suicide victims be our population.

Solution:

a. $H_0 : p_{\text{W}} = p_{\text{B}}$
b. $H_a : p_{\text{W}} \neq p_{\text{B}}$
c. The random variable is the difference in the sample proportions of white and black suicide victims, aged 15 to 24.
d. normal for two proportions
e. test statistic: $z = –0.1944$
f. p-value: 0.8458
g. Check student’s solution.

h.
   i. Alpha: 0.05
   ii. Decision: Reject the null hypothesis.
   iii. Reason for decision: p-value > alpha
   iv. Conclusion: At the 5% significance level, there is insufficient evidence to conclude that the proportions of

white and black female suicide victims, aged 15 to 24, are different.

Exercise:

Problem:

Elizabeth Mjelde, an art history professor, was interested in whether the value from the Golden Ratio formula, $\left(\frac{\text{larger + smaller dimension}}{\text{larger dimension}}\right)$ was the same in the Whitney Exhibit for works from 1900 to 1919 as for works from 1920 to 1942. Thirty-seven early works were sampled, averaging 1.74 with a standard deviation of 0.11. Sixty-five of the later works were sampled, averaging 1.746 with a standard deviation of 0.1064. Do you think that there is a significant difference in the Golden Ratio calculation?

Exercise:

Problem:

A recent year was randomly picked from 1985 to the present. In that year, there were 2,051 Hispanic students at Cabrillo College out of a total of 12,328 students. At Lake Tahoe College, there were 321 Hispanic students out of a total of 2,441 students. In general, do you think that the percent of Hispanic students at the two colleges is basically the same or different?

Solution:

Subscripts: 1 = Cabrillo College, 2 = Lake Tahoe College

a. $H_0 : p_1 = p_2$
b. $H_a : p_1 \neq p_2$
c. The random variable is the difference between the proportions of Hispanic students at Cabrillo College and Lake Tahoe College.
d. normal for two proportions
e. test statistic: $z = 4.29$
f. p-value: 0.00002

g. Check student’s solution.

h.
  i. Alpha: 0.05
  ii. Decision: Reject the null hypothesis.
  iii. Reason for decision: p-value < alpha
  iv. Conclusion: There is sufficient evidence to conclude that the proportions of Hispanic students at Cabrillo College and Lake Tahoe College are different.

Use the following information to answer the next three exercises. Neuroinvasive West Nile virus is a severe disease that affects a person’s nervous system . It is spread by the Culex species of mosquito. In the United States in 2010 there were 629 reported cases of neuroinvasive West Nile virus out of a total of 1,021 reported cases and there were 486 neuroinvasive reported cases out of a total of 712 cases reported in 2011. Is the 2011 proportion of neuroinvasive West Nile virus cases more than the 2010 proportion of neuroinvasive West Nile virus cases? Using a 1% level of significance, conduct an appropriate hypothesis test.

- “2011” subscript: 2011 group.
- “2010” subscript: 2010 group

Exercise:

Problem: This is:

a. a test of two proportions
b. a test of two independent means
c. a test of a single mean
d. a test of matched pairs.

Exercise:

Problem: An appropriate null hypothesis is:

a. $p_{2011} = p_{2010}$

b. $p_{2011} \geq p_{2010}$
c. $\mu_{2011} = \mu_{2010}$
d. $p_{2011} > p_{2010}$

Solution:

a

Exercise:

Problem:

The p-value is 0.0022. At a 1% level of significance, the appropriate conclusion is

a. There is sufficient evidence to conclude that the proportion of people in the United States in 2011 who contracted neuroinvasive West Nile disease is less than the proportion of people in the United States in 2010 who contracted neuroinvasive West Nile disease.
b. There is insufficient evidence to conclude that the proportion of people in the United States in 2011 who contracted neuroinvasive West Nile disease is more than the proportion of people in the United States in 2010 who contracted neuroinvasive West Nile disease.
c. There is insufficient evidence to conclude that the proportion of people in the United States in 2011 who contracted neuroinvasive West Nile disease is less than the proportion of people in the United States in 2010 who contracted neuroinvasive West Nile disease.
d. There is sufficient evidence to conclude that the proportion of people in the United States in 2011 who contracted neuroinvasive West Nile disease is more than the proportion of people in the United States in 2010 who contracted neuroinvasive West Nile disease.

Exercise:

Problem:

Researchers conducted a study to find out if there is a difference in the use of eReaders by different age groups. Randomly selected participants were divided into two age groups. In the 16- to 29-year-old group, 7% of the 628 surveyed use eReaders, while 11% of the 2,309 participants 30 years old and older use eReaders. Run an appropriate hypothesis test.

Solution:

Random variable: $\widehat{P}_1 - \widehat{P}_2$, where subscripts 1 and 2 represent the age groups of 16-29 years and 30 years and older, respectively.

Hypotheses:
$H_0 : p_1 = p_2$
$H_a : p_1 \neq p_2$

p-value : 0.0033

Decision: Reject the null hypothesis.

Conclusion: At the 5% level of significance, from the sample data, there is sufficient evidence to conclude that the proportion of eReader users ages 16 to 29 years old is different from the proportion of eReader users that are 30 and older.

Exercise:

Problem:

Adults aged 18 years old and older were randomly selected for a survey on obesity. Adults are considered obese if their body mass index (BMI) is at least 30. The researchers wanted to determine if the proportion of women who are obese in the south is less than the proportion of southern men who are obese. The results are shown in the table below. Test at the 1% level of significance.

	Number who are obese	**Sample size**
Men	42,769	155,525
Women	67,169	248,775

Exercise:

Problem:

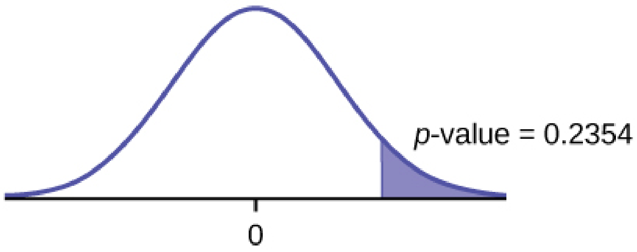
Two computer users were discussing tablet computers. We are interested to know whether a higher proportion of people ages 16 to 29 use tablets than the proportion of people age 30 and older. The table below details the number of tablet owners for each age group. Test at the 1% level of significance.

	16–29 year olds	**30 years old and older**
Own a Tablet	69	231
Sample Size	628	2,309

Solution:

Random variable: $\widehat{P}_1 - \widehat{P}_2$, where subscripts 1 and 2 represent the age groups of 16-29 years and 30 years and older, respectively.

Hypotheses:
$H_0 : p_1 = p_2$
$H_a : p_1 > p_2$



p-value: 0.2354

Decision: Do not reject the H_0 .

Conclusion: At the 1% level of significance, from the sample data, there is not sufficient evidence to conclude that a higher proportion of tablet owners are aged 16 to 29 years old than are 30 years old and older.

Exercise:

Problem:

A group of friends debated whether more men use smartphones than women. They consulted a research study of smartphone use among adults. The results of the survey indicate that of the 973 men randomly sampled, 379 use smartphones. For women, 404 of the 1,304 who were randomly sampled use smartphones. Test at the 5% level of significance.

Exercise:

Problem:

While her husband spent 2½ hours picking out new speakers, a statistician decided to determine whether the percent of men who enjoy shopping for electronic equipment is higher than the percent of women who enjoy shopping for electronic equipment. The population was Saturday afternoon shoppers. Out of 67 men, 24 said they enjoyed the activity. Eight of the 24 women surveyed claimed to enjoy the activity. Interpret the results of the survey.

Solution:

Subscripts: 1: men; 2: women

a. $H_0 : p_1 = p_2$
b. $H_a : p_1 > p_2$
c. $\widehat{P}_1 - \widehat{P}_2$ is the difference between the proportions of men and women who enjoy shopping for electronic equipment.
d. normal for two proportions
e. test statistic: $z = 0.22$
f. p-value: 0.4133
g. Check student's solution.

h.
   - i. Alpha: 0.05
   - ii. Decision: Do not reject the null hypothesis.
   - iii. Reason for Decision: p-value > alpha
   - iv. Conclusion: At the 5% significance level, there is insufficient evidence to conclude that the proportion of men who enjoy shopping for electronic equipment is more than the proportion of women.

i. Note: Be cautious of these results. Notice that in one of the samples, the number of successes is less than 10, which means a normal distribution may not be appropriate. Thus, the results may be invalid.

Exercise:

Problem:

We are interested in whether children's educational computer software costs less, on average, than children's entertainment software. Thirty-six educational software titles were randomly picked from a catalog. The mean cost was $31.14 with a standard deviation of $4.69. Thirty-five entertainment software titles were randomly picked from the same catalog. The mean cost was $33.86 with a standard deviation of $10.87. Decide whether children's educational software costs less, on average, than children's entertainment software.

Exercise:

Problem:

Joan Nguyen recently claimed that the proportion of college-age males with at least one pierced ear is as high as the proportion of college-age females. She conducted a survey in her classes. Out of 107 males, 20 had at least one pierced ear. Out of 92 females, 47 had at least one pierced ear. Do you believe that the proportion of college-age males with at least one pierced ear has reached the proportion of females?

Solution:

a. $H_0 : p_1 = p_2$
b. $H_a : p_1 \neq p_2$
c. $\widehat{P}_1 - \widehat{P}_2$ is the difference between the proportions of men and women that have at least one pierced ear.
d. normal for two proportions
e. test statistic: $z = –4.82$
f. p-value: 0.000001
g. Check student’s solution.

h.
   i. Alpha: 0.05
   ii. Decision: Reject the null hypothesis.
   iii. Reason for Decision: p-value < alpha
   iv. Conclusion: At the 5% significance level, there is sufficient evidence to conclude that the proportions of males and females with at least one pierced ear is different.

Exercise:

Problem:

Use the data sets found in Appendix C to answer this exercise. Is the proportion of race laps Terri completes slower than 130 seconds less than the proportion of practice laps she completes slower than 135 seconds?

Glossary

Pooled Proportion
estimate of the common value of p_1 and p_2.

Matched or Paired Samples

When using a hypothesis test for **matched or paired samples**, the following characteristics should be present:

1. Simple random sampling is used.
2. Sample sizes are often small.
3. Two measurements are drawn from the each pair of individuals or objects.
4. Differences are calculated from the matched or paired samples.
5. The differences form the sample that is used for the hypothesis test.
6. Either the matched pairs have differences that come from a population that is normal or the number of differences is sufficiently large so that distribution of the sample mean of differences is approximately normal.

Recall that **matched or paired samples** are samples in which individuals from each sample have been matched or paired in some meaningful way. This may include individuals who've been paired with themselves.

In a hypothesis test for matched or paired samples, subjects are matched in pairs and differences are calculated. **The differences are the data.** The population mean for the differences, μ_d, is then tested using a Student's t-test for a single population mean with $n – 1$ degrees of freedom, where n is the number of differences.

The competing hypotheses will have one of the following forms:

$H_0 : \mu_\text{d} = 0; H_a : \mu_\text{d} \neq 0$

$H_0 : \mu_\text{d} = 0; H_a : \mu_\text{d} > 0$

$H_0 : \mu_\text{d} = 0; H_a : \mu_\text{d} < 0$

Equation:

The test statistic (*t*-score) is:

$$t = \frac{x_d - \mu_d}{\left(\frac{s_d}{\sqrt{n}}\right)}$$

Example:

Exercise:

Problem:

A study was conducted to investigate the effectiveness of hypnotism in reducing pain. Results for randomly selected subjects are shown in the following table. A lower score indicates less pain. The "before" value is matched to an "after" value and the differences are calculated. The differences have a normal distribution. Are the sensory measurements, on average, lower after hypnotism? Test at a 5% significance level.

Subject:	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**
Before	6.6	6.5	9.0	10.3	11.3	8.1	6.3	11.6
After	6.8	2.4	7.4	8.5	8.1	6.1	3.4	2.0

Solution:

Corresponding "before" and "after" values form matched pairs. (Calculate "after" – "before.")

After Data	**Before Data**	**Difference**
6.8	6.6	0.2
2.4	6.5	-4.1
7.4	9	-1.6
8.5	10.3	-1.8
8.1	11.3	-3.2
6.1	8.1	-2
3.4	6.3	-2.9
2	11.6	-9.6

The data for the test are the differences: {0.2, –4.1, –1.6, –1.8, –3.2, –2, –2.9, –9.6}

The sample mean and sample standard deviation of the differences are: $x_d =$ –3.13 and $s_d =$ 2.91. Verify these values.

Let μ_d be the population mean for the differences. We use the subscript d to denote "differences."

Random variable: X_d = the mean difference of the sensory measurements.

H_0 : μ_d = 0

The null hypothesis is that the difference is zero, meaning that there is the same amount of pain felt after hypnotism. That means the subject shows no improvement. (μ_d is the population mean of the differences.)

H_a : $\mu_{\text{d}} < 0$

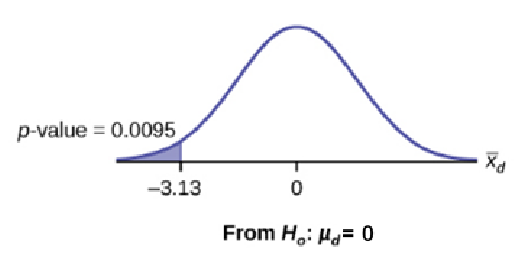
The alternative hypothesis is negative, meaning there is less pain felt after hypnotism. That means the subject shows improvement. The score should be lower after hypnotism, so the difference ought to be negative to indicate improvement.

Distribution for the test: The distribution is a Student's t with $df = n – 1 = 8 – 1 = 7$. Use t_7.

(Notice that the test is for a single population mean.)

Calculate the p-value using the Student's t-distribution: p-value = 0.0095

Graph:



X_d is the random variable for the differences.

The sample mean and sample standard deviation of the differences are:

x_d = –3.13

s_d = 2.91

Compare α and the p-value: α = 0.05 and p-value = 0.0095. p-value < α.

Make a decision: Since p-value < α, reject H_0 . This means that $\mu_{\text{d}} < 0$ and there is improvement.

Conclusion: At a 5% level of significance, from the sample data, there is sufficient evidence to conclude that the sensory measurements, on average, are lower after hypnotism. Hypnotism appears to be effective in reducing pain.

Note:
Note
For the TI-83+ and TI-84 calculators, you can either calculate the differences ahead of time (**after - before**) and put the differences into a list or you can put the **after** data into a first list and the **before** data into a second list. Then go to a third list and arrow up to the name. Enter 1st list name - 2nd list name. The calculator will do the subtraction, and you will have the differences in the third list.

Note:
Use your list of differences as the data. Press `STAT` and arrow over to `TESTS`. Press `2:T-Test`. Arrow over to `Data` and press `ENTER`. Arrow down and enter `0` for μ_0, the name of the list where you put the data, and `1` for Freq:. Arrow down to `μ`: and arrow over to `<` μ_0. Press `ENTER`. Arrow down to `Calculate` and press `ENTER`. The *p*-value is 0.0094, and the test statistic is -3.04. Do these instructions again except, arrow to `Draw` (instead of `Calculate`). Press `ENTER`.

Note:
Try It
Exercise:

Problem:

A study was conducted to investigate how effective a new diet was in lowering cholesterol. Results for the randomly selected subjects are shown in the table. The differences have a normal distribution. Are the subjects' cholesterol levels lower on average after the diet? Test at the 5% level.

Subject	A	B	C	D	E	F	G	H	I
Before	209	210	205	198	216	217	238	240	222
After	199	207	189	209	217	202	211	223	201

Solution:

The *p*-value is 0.0130, so we can reject the null hypothesis. There is enough evidence to suggest that the diet lowers cholesterol.

Example:
A college football coach was interested in whether the college's strength development class increased his players' maximum lift (in pounds) on the bench press exercise. He asked four of his players to participate in a study. The amount of weight they could each lift was recorded before they took the strength development

class. After completing the class, the amount of weight they could each lift was again measured. The data are as follows:

Weight (in pounds)	**Player 1**	**Player 2**	**Player 3**	**Player 4**
Amount of weight lifted prior to the class	205	241	338	368
Amount of weight lifted after the class	295	252	330	360

The coach wants to know if the strength development class makes his players stronger, on average.

Record the **differences** data. Calculate the differences by subtracting the amount of weight lifted prior to the class from the weight lifted after completing the class. The data for the differences are: {90, 11, -8, -8}. Assume the differences have a normal distribution.

Using the differences data, calculate the sample mean and the sample standard deviation.

x_d = 21.3, s_{d} = 46.7

Note:
Note
The data given here would indicate that the distribution is actually right-skewed. The difference 90 may be an extreme outlier? It is pulling the sample mean to be 21.3 (positive). The means of the other three data values are actually negative.

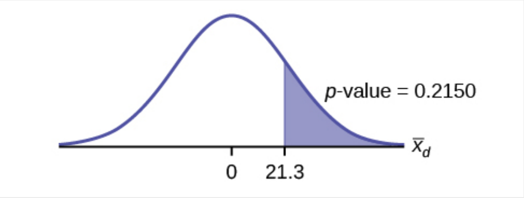
Using the difference data, this becomes a test of a single __________ (fill in the blank).

Define the random variable: X_d mean difference in the maximum lift per player.

The distribution for the hypothesis test is t_3.

$H_0 : \mu_{\text{d}} = 0$; $H_a : \mu_{\text{d}} > 0$

Graph:



Calculate the p-value: The p-value is 0.2150

Decision: If the level of significance is 5%, the decision is not to reject the null hypothesis, because the p-value > α.

What is the conclusion?

At a 5% level of significance, from the sample data, there is not sufficient evidence to conclude that the strength development class helped to make the players stronger, on average.

Note:
Try It
Exercise:

Problem:

A new prep class was designed to improve SAT test scores. Four students were selected at random. Their scores on two practice exams were recorded, one before the class and one after. The data recorded in the following table. Are the scores, on average, higher after the class? Test at a 5% level.

SAT Scores	**Student 1**	**Student 2**	**Student 3**	**Student 4**
Score before class	1840	1960	1920	2150
Score after class	1920	2160	2200	2100

Solution:

The *p*-value is 0.0874, so we decline to reject the null hypothesis. The data do not support that the class improves SAT scores significantly.

Example:
Seven eighth graders at Kennedy Middle School measured how far they could push the shot-put with their dominant (writing) hand and their weaker (non-writing) hand. They thought that they could push equal distances with either hand. The data were collected and recorded in the following table.

Distance (in feet) using	**Student 1**	**Student 2**	**Student 3**	**Student 4**	**Student 5**	**Student 6**	**Student 7**
Dominant Hand	30	26	34	17	19	26	20

Distance (in feet) using	**Student 1**	**Student 2**	**Student 3**	**Student 4**	**Student 5**	**Student 6**	**Student 7**
Weaker Hand	28	14	27	18	17	26	16

Conduct a hypothesis test to determine whether the mean difference in distances between the children's dominant versus weaker hands is significant.

Record the **differences** data. Calculate the differences by subtracting the distances with the weaker hand from the distances with the dominant hand. The data for the differences are: {2, 12, 7, –1, 2, 0, 4}. The differences have a normal distribution.

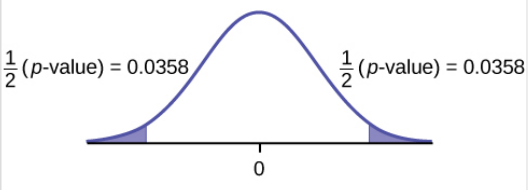
Using the differences data, calculate the sample mean and the sample standard deviation.

x_d = 3.71, s_d = 4.5.

Random variable: X_d = mean difference in the distances between the hands.

Distribution for the hypothesis test: t_6.

$H_0 : \mu_\text{d} = 0$; $H_a : \mu_\text{d} \neq 0$



Calculate the p-value: The p-value is 0.0716 (using the data directly).

test statistic = 2.18. p-value = 0.0719 using $\left(x_d = 3.71, s_d = 4.5.\right)$

Decision: Assume α = 0.05. Since the p-value > α, do not reject H_0 .

Conclusion: At the 5% level of significance, from the sample data, there is not sufficient evidence to conclude that there is a difference in the children's weaker and dominant hands to push the shot-put.

Note:
Try-It
Exercise:

Problem:

Five ball players think they can throw the same distance with their dominant hand (throwing) and off-hand (catching hand). The data were collected and recorded in the following table. Conduct a hypothesis test to determine whether the mean difference in distances between the dominant and off-hand is significant. Test at the 5% level.

	Player 1	Player 2	Player 3	Player 4	Player 5
Dominant Hand	120	111	135	140	125
Off-hand	105	109	98	111	99

Solution:

The p-value is 0.0230, so we can reject the null hypothesis. The data show that the players do not throw the same distance with their off-hands as they do with their dominant hands.

Section Review

A hypothesis test for matched or paired samples (t-test) has these characteristics:

- Test the differences by subtracting one measurement from the other measurement
- Random Variable: x_d = mean of the differences
- Distribution: Student's t-distribution with $n – 1$ degrees of freedom
- If the number of differences is small (less than 30), the differences must follow a normal distribution.
- Two measurements are drawn from each pair of individuals or objects.
- Samples are dependent, that is samples are matched or paired.

Formula Review

Test Statistic (t-score): $t = \frac{x_d - \mu_d}{\left(\frac{s_d}{\sqrt{n}}\right)}$

where x_d is the mean of the sample differences. μ_{d} is the mean of the population differences. s_{d} is the sample standard deviation of the differences. n is the sample size.

Use the following information to answer the next five exercises. A study was conducted to test the effectiveness of a software patch in reducing system failures over a six-month period. Results for randomly selected installations are shown in the following table. The “before” value is matched to an “after” value, and the differences are calculated. The differences have a normal distribution. Test at the 1% significance level.

Installation	A	B	C	D	E	F	G	H
Before	3	6	4	2	5	8	2	6
After	1	5	2	0	1	0	2	2

Exercise:

Problem: What is the random variable?

Solution:

the mean difference of the system failures

Exercise:

Problem: State the null and alternative hypotheses.

Exercise:

Problem: What is the p-value?

Solution:

0.0067

Exercise:

Problem: Draw the graph of the p-value.

Exercise:

Problem: What conclusion can you draw about the software patch?

Solution:

With a p-value 0.0067, we can reject the null hypothesis. There is enough evidence to support that the software patch is effective in reducing the number of system failures.

Use the following information to answer next five exercises. A study was conducted to test the effectiveness of a juggling class. Before the class started, six subjects juggled as many balls as they could at once. After the class, the same six subjects juggled as many balls as they could. The differences in the number of balls are calculated. The differences have a normal distribution. Test at the 1% significance level.

Subject	**A**	**B**	**C**	**D**	**E**	**F**
Before	3	4	3	2	4	5
After	4	5	6	4	5	7

Exercise:

Problem: State the null and alternative hypotheses.

Exercise:

Problem: What is the p-value?

Solution:

0.002

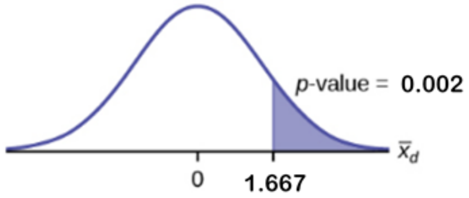
Exercise:

Problem: What is the sample mean difference?

Exercise:

Problem: Draw the graph of the *p*-value.

Solution:



Exercise:

Problem: What conclusion can you draw about the juggling class?

Use the following information to answer the next five exercises. A doctor wants to know if a blood pressure medication is effective. Six subjects have their blood pressures recorded. After twelve weeks on the medication, the same six subjects have their blood pressure recorded again. For this test, only systolic pressure is of concern. Test at the 1% significance level.

Patient	**A**	**B**	**C**	**D**	**E**	**F**
Before	161	162	165	162	166	171
After	158	159	166	160	167	169

Exercise:

Problem: State the null and alternative hypotheses.

Solution:

$H_0 : \mu_d = 0$

$H_a : \mu_d < 0$

Exercise:

Problem: What is the test statistic?

Exercise:

Problem: What is the p-value?

Solution:

0.0699

Exercise:

Problem: What is the sample mean difference?

Exercise:

Problem: What is the conclusion?

Solution:

We decline to reject the null hypothesis. There is not sufficient evidence to support that the medication is effective.

Homework

For each of the word problems below, use a solution sheet to do the hypothesis test. The Solutions Sheets can be found in the Table of Contents or by clicking *here*. *Please feel free to make copies of the Solution Sheets. For the online version of the book, it is suggested that you copy the .doc or the .pdf files.*

Note:
Note
If you are using a Student's t-distribution for the homework problems, including for paired data, you may assume that the underlying population is normally distributed. (When using these tests in a real situation, you must first prove that assumption, however.)

Exercise:

Problem:

Ten individuals went on a low–fat diet for 12 weeks to lower their cholesterol. The data are recorded in the following table. Do you think that their cholesterol levels were significantly lowered?

Starting cholesterol level	Ending cholesterol level
140	140
220	230

Starting cholesterol level	**Ending cholesterol level**
110	120
240	220
200	190
180	150
190	200
360	300
280	300
260	240

Solution:

p-value = 0.1494

At the 5% significance level, there is insufficient evidence to conclude that the medication lowered cholesterol levels after 12 weeks.

Use the following information to answer the next two exercises. A new AIDS prevention drug was tried on a group of 224 HIV positive patients. Forty-five patients developed AIDS after four years. In a control group of 224 HIV positive patients, 68 developed AIDS after four years. We want to test whether the method of treatment reduces the proportion of patients that develop AIDS after four years or if the proportions of the treated group and the untreated group stay the same.

Let the subscript t = treated patient and u = untreated patient.

Exercise:

Problem: The appropriate hypotheses are:

a. $H_0 : p_t < p_u$ and $H_a : p_t = p_u$
b. $H_0 : p_t \leq p_u$ and $H_a : p_t > p_u$
c. $H_0 : p_t = p_u$ and $H_a : p_t \neq p_u$
d. $H_0 : p_t = p_u$ and $H_a : p_t < p_u$

Exercise:

Problem: If the p-value is 0.0062 what is the conclusion (use α = 0.05)?

a. The method has no effect.
b. There is sufficient evidence to conclude that the method reduces the proportion of HIV positive patients who develop AIDS after four years.
c. There is sufficient evidence to conclude that the method increases the proportion of HIV positive patients who develop AIDS after four years.

d. There is insufficient evidence to conclude that the method reduces the proportion of HIV positive patients who develop AIDS after four years.

Solution:

b

Use the following information to answer the next two exercises. An experiment is conducted to show that blood pressure can be consciously reduced in people trained in a “biofeedback exercise program.” Six subjects were randomly selected and blood pressure measurements were recorded before and after the training. The difference between blood pressures was calculated (after - before) producing the following results: $x_d = -10.2$ $s_{\text{d}} = 8.4$. Using the data, test the hypothesis that the blood pressure has decreased after the training.

Exercise:

Problem: The distribution for the test is:

a. t_5
b. t_6
c. $N(-10.2, 8.4)$
d. $N(-10.2, \frac{8.4}{\sqrt{6}})$

Exercise:

Problem: If $\alpha = 0.05$, the p-value and the conclusion are

a. 0.0014; There is sufficient evidence to conclude that the blood pressure decreased after the training.
b. 0.0014; There is sufficient evidence to conclude that the blood pressure increased after the training.
c. 0.0155; There is sufficient evidence to conclude that the blood pressure decreased after the training.
d. 0.0155; There is sufficient evidence to conclude that the blood pressure increased after the training.

Solution:

c

Exercise:

Problem:

A golf instructor is interested in determining if her new technique for improving players’ golf scores is effective. She takes four new students. She records their 18-hole scores before learning the technique and then after having taken her class. She conducts a hypothesis test. The data are as follows.

	Player 1	**Player 2**	**Player 3**	**Player 4**
Mean score before class	83	78	93	87
Mean score after class	80	80	86	86

The correct decision is:

a. Reject H_0.
b. Do not reject the H_0.

Exercise:

Problem:

A local cancer support group believes that the estimate for new female breast cancer cases in the south is higher in 2013 than in 2012. The group compared the estimates of new female breast cancer cases by southern state in 2012 and in 2013. The results are in the following table.

Southern States	**2012**	**2013**
Alabama	3,450	3,720
Arkansas	2,150	2,280
Florida	15,540	15,710
Georgia	6,970	7,310
Kentucky	3,160	3,300
Louisiana	3,320	3,630
Mississippi	1,990	2,080
North Carolina	7,090	7,430
Oklahoma	2,630	2,690
South Carolina	3,570	3,580
Tennessee	4,680	5,070
Texas	15,050	14,980
Virginia	6,190	6,280

Solution:

Test: two matched pairs or paired samples (t-test)

Random variable: X_d

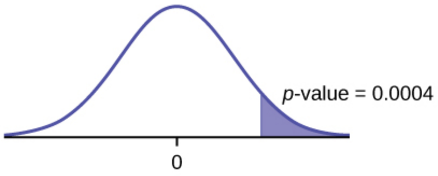
Distribution: t_{12}

$H_0 : \mu_d = 0$ $H_a : \mu_d > 0$

The mean of the differences of new female breast cancer cases in the south between 2013 and 2012 is greater than zero. The estimate for new female breast cancer cases in the south is higher in 2013 than in 2012.

Graph: right-tailed

p-value: 0.0004



Decision: Reject H_0

Conclusion: At the 5% level of significance, from the sample data, there is sufficient evidence to conclude that there was a higher estimate of new female breast cancer cases in 2013 than in 2012.

Exercise:

Problem:

A traveler wanted to know if the prices of hotels are different in the ten cities that he visits the most often. The list of the cities with the corresponding hotel prices for his two favorite hotel chains is in the following table. Test at the 1% level of significance.

Cities	**Hyatt Regency prices in dollars**	**Hilton prices in dollars**
Atlanta	107	169
Boston	358	289
Chicago	209	299
Dallas	209	198
Denver	167	169
Indianapolis	179	214
Los Angeles	179	169
New York City	625	459
Philadelphia	179	159
Washington, DC	245	239

Exercise:

Problem:

A politician asked his staff to determine whether the underemployment rate in the northeast decreased from 2011 to 2012. The results are in the following table.

Northeastern States	**2011**	**2012**
Connecticut	17.3	16.4
Delaware	17.4	13.7
Maine	19.3	16.1
Maryland	16.0	15.5
Massachusetts	17.6	18.2
New Hampshire	15.4	13.5
New Jersey	19.2	18.7
New York	18.5	18.7
Ohio	18.2	18.8
Pennsylvania	16.5	16.9
Rhode Island	20.7	22.4
Vermont	14.7	12.3
West Virginia	15.5	17.3

Solution:

Test: matched or paired samples (t-test)

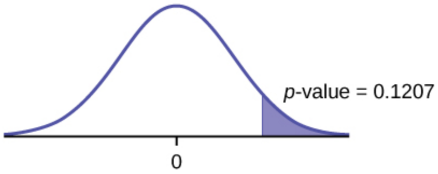
Difference data: {0.9, 3.7, 3.2, 0.5, –0.6, 1.9, 0.5, –0.2, –0.6, –0.4, –1.7, 2.4, –1.8}

Random Variable: X_d

Distribution: $H_0 : \mu_{\mathrm{d}} = 0$; $H_a : \mu_{\mathrm{d}} > 0$

The mean of the differences of the rate of underemployment in the northeastern states between 2011 and 2012 (2011 values minus 2012 values) is greater than zero. The underemployment rate went down from 2011 to 2012.

Graph: right-tailed.



p-value: 0.1207

Decision: Do not reject H_0.

Conclusion: At the 5% level of significance, from the sample data, there is not sufficient evidence to conclude that there was a decrease in the underemployment rates of the northeastern states from 2011 to 2012.

Exercise:

Problem: "To Breakfast or Not to Breakfast?" by Richard Ayore

In the American society, birthdays are one of those days that everyone looks forward to. People of different ages and peer groups gather to mark the 18th, 20th, …, birthdays. During this time, one looks back to see what he or she has achieved for the past year and also focuses ahead for more to come.

If, by any chance, I am invited to one of these parties, my experience is always different. Instead of dancing around with my friends while the music is booming, I get carried away by memories of my family back home in Kenya. I remember the good times I had with my brothers and sister while we did our daily routine.

Every morning, I remember we went to the shamba (garden) to weed our crops. I remember one day arguing with my brother as to why he always remained behind just to join us an hour later. In his defense, he said that he preferred waiting for breakfast before he came to weed. He said, “This is why I always work more hours than you guys!”

And so, to prove him wrong or right, we decided to give it a try. One day we went to work as usual without breakfast, and recorded the time we could work before getting tired and stopping. On the next day, we all ate breakfast before going to work. We recorded how long we worked again before getting tired and stopping. Of interest was our mean increase in work time. Though not sure, my brother insisted that it was more than two hours. Using the data in the following table, solve our problem.

Work hours with breakfast	**Work hours without breakfast**
8	6
7	5
9	5
5	4

Work hours with breakfast	**Work hours without breakfast**
9	7
8	7
10	7
7	5
6	6
9	5

Solution:

a. $H_0 : \mu_d = 0$
b. $H_a : \mu_d > 0$
c. The random variable X_d is the mean difference in work times on days when eating breakfast and on days when not eating breakfast.
d. t_9
e. test statistic: $t = 4.8963$
f. p-value: 0.0004
g. Check student's solution.
h.
    i. Alpha: 0.05
    ii. Decision: Reject the null hypothesis.
    iii. Reason for Decision: p-value < alpha
    iv. Conclusion: At the 5% level of significance, there is sufficient evidence to conclude that the mean difference in work times on days when eating breakfast and on days when not eating breakfast has increased.

Bringing It Together

Use the following information to answer the next ten exercises. indicate which of the following choices best identifies the hypothesis test.

a. independent group means, population standard deviations unknown
b. matched or paired samples
c. single mean
d. two proportions
e. single proportion

Exercise:

Problem:

A powder diet is tested on 49 people, and a liquid diet is tested on 36 different people. The population standard deviations are two pounds and three pounds, respectively. Of interest is whether the liquid diet yields a higher mean weight loss than the powder diet.

Exercise:

Problem:

A new chocolate bar is taste-tested on consumers. Of interest is whether the proportion of children who like the new chocolate bar is greater than the proportion of adults who like it.

Solution:

d

Exercise:

Problem:

The mean number of English courses taken in a two–year time period by male and female college students is believed to be about the same. An experiment is conducted and data are collected from nine males and 16 females.

Exercise:

Problem:

A football league reported that the mean number of touchdowns per game was five. A study is done to determine if the mean number of touchdowns has decreased.

Solution:

c

Exercise:

Problem:

A study is done to determine if students in the California state university system take longer to graduate than students enrolled in private universities. One hundred students from both the California state university system and private universities are surveyed. From years of research, it is known that the population standard deviations are 1.5811 years and one year, respectively.

Exercise:

Problem:

According to a YWCA Rape Crisis Center newsletter, 75% of rape victims know their attackers. A study is done to verify this.

Solution:

e

Exercise:

Problem: According to a recent study, U.S. companies have a mean maternity-leave of six weeks.

Exercise:

Problem:

A recent drug survey showed an increase in use of drugs and alcohol among local high school students as compared to the national percent. Suppose that a survey of 100 local youths and 100 national youths is conducted to see if the proportion of drug and alcohol use is higher locally than nationally.

Solution:

d

Exercise:

Problem:

A new SAT study course is tested on 12 individuals. Pre-course and post-course scores are recorded. Of interest is the mean increase in SAT scores. The following data are collected:

Pre-course score	**Post-course score**
1	300
960	920
1010	1100
840	880
1100	1070
1250	1320
860	860
1330	1370
790	770
990	1040
1110	1200
740	850

Exercise:

Problem:

University of Michigan researchers reported in the *Journal of the National Cancer Institute* that quitting smoking is especially beneficial for those under age 49. In this American Cancer Society study, the risk (probability) of dying of lung cancer was about the same as for those who had never smoked.

Solution:

e

Exercise:

Problem:

Lesley E. Tan investigated the relationship between left-handedness vs. right-handedness and motor competence in preschool children. Random samples of 41 left-handed preschool children and 41 right-handed preschool children were given several tests of motor skills to determine if there is evidence of a difference between the children based on this experiment. The experiment produced the means and standard deviations shown in the following table. Determine the appropriate test and best distribution to use for that test.

	Left-handed	Right-handed
Sample size	41	41
Sample mean	97.5	98.1
Sample standard deviation	17.5	19.2

a. Two independent means, normal distribution
b. Two independent means, Student’s t-distribution
c. Matched or paired samples, Student’s t-distribution
d. Two population proportions, normal distribution

Exercise:

Problem:

A golf instructor is interested in determining if her new technique for improving players’ golf scores is effective. She takes four (4) new students. She records their 18-hole scores before learning the technique and then after having taken her class. She conducts a hypothesis test. The data are as follows.

	Player 1	**Player 2**	**Player 3**	**Player 4**
Mean score before class	83	78	93	87
Mean score after class	80	80	86	86

This is:

a. a test of two independent means.
b. a test of two proportions.
c. a test of a single mean.
d. a test of a single proportion.

Solution:

a

Glossary

matched or paired samples	samples in which individuals from each sample have been matched or paired in some meaningful way. This may include individuals who've been paired with themselves.

The Chi Square Distribution: Introduction class="introduction"

The chi-square distribution can be used to find relationships between two things, like grocery prices at different stores. (credit: Pete/flickr)

Note:
Chapter Objectives
By the end of this chapter, the student should be able to:

- Interpret the chi-square probability distribution as the sample size changes.
- Conduct and interpret chi-square goodness-of-fit hypothesis tests.
- Conduct and interpret chi-square test of independence hypothesis tests.

Have you ever wondered if lottery numbers were evenly distributed or if some numbers occurred with a greater frequency? How about if the types of movies people preferred were different across different age groups? What about if a coffee machine was dispensing approximately the same amount of coffee each time? You could answer these questions by conducting a hypothesis test.

You will now study a new distribution, one that is used to determine the answers to such questions. This distribution is called the chi-square distribution.

In this chapter, you will learn the two major applications of the chi-square distribution:

1. the goodness-of-fit test, which determines if data fit a particular distribution, such as in the lottery example
2. the test of independence, which determines if events are independent, such as in the movie example

Note:
NOTE

Though the chi-square distribution depends on calculators or computers for most of the calculations, there is a table available (see Appendix B). TI-83+ and TI-84 calculator instructions are included in the text.

Note:

Collaborative Classroom Exercise

Look in the sports section of a newspaper or on the Internet for some sports data (baseball averages, basketball scores, golf tournament scores, football odds, swimming times, and the like). Plot a histogram and a boxplot using your data. See if you can determine a probability distribution that your data fits. Have a discussion with the class about your choice.

Facts About the Chi-Square Distribution

The notation for the **chi-square distribution** is:
Equation:

$$X \sim \chi^2_{df}$$

where df = degrees of freedom, which depends on how chi-square is being used. (If you want to practice calculating chi-square probabilities then use $df = n - 1$. The degrees of freedom for the two major uses are each calculated differently.)

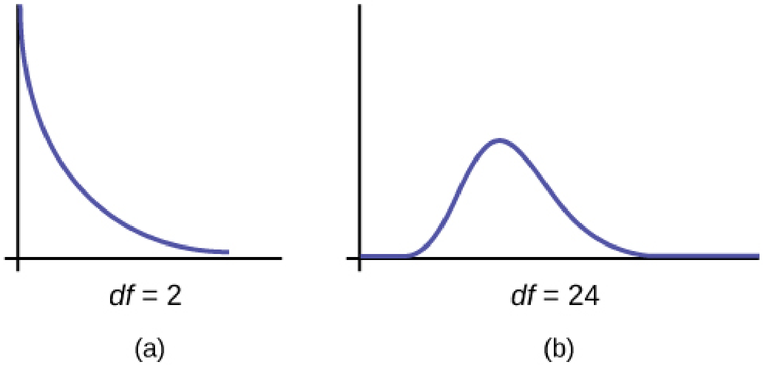
For the χ^2 distribution, the population mean is $\mu = df$ and the population standard deviation is $\sigma = \sqrt{2(df)}$.

The random variable is shown as χ^2, but may be any upper case letter.

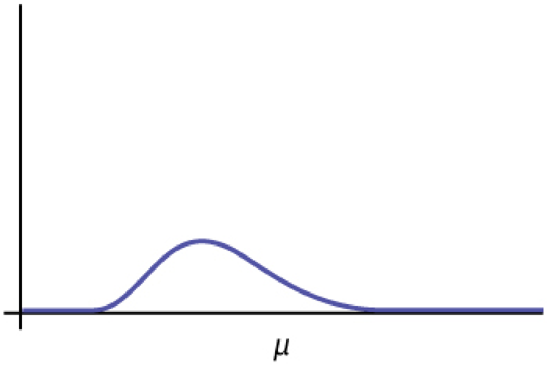
The random variable for a chi-square distribution with k degrees of freedom is the sum of k independent, squared standard normal variables.

$\chi^2 = (Z_1)^2 + (Z_2)^2 + ... + (Z_k)^2$

1. The curve is nonsymmetrical and skewed to the right.
2. There is a different chi-square curve for each df.



3. The test statistic for any test is always greater than or equal to zero.
4. When $df > 90$, the chi-square curve approximates the normal distribution. For $X \sim \chi^2_{1,000}$ the mean, $\mu = df = 1{,}000$ and the standard deviation, $\sigma = \sqrt{2(1{,}000)} = 44.7$. Therefore, $X \sim N(1{,}000,\ 44.7)$, approximately.
5. The mean, μ, is located just to the right of the peak.



Note: Another interesting fact is that "the Poisson distribution resembles a discrete version of the chi-square distribution".[footnote] Recall that the Poisson Distribution was covered in Chapter 4.
Chi-Squared Distribution by Michael Manser, Subhiskha Swamy, and James Blanchard (http://www.colorado.edu/economics/morey/7818/univariatervs/Chi-squared/ECON%207818%20Chi.pdf

References

Data from *Parade Magazine*.

“HIV/AIDS Epidemiology Santa Clara County.”Santa Clara County Public Health Department, May 2011.

Section Review

The chi-square distribution is a useful tool for assessment in a series of problem categories. These problem categories include primarily (i) whether a data set fits a particular distribution, (ii) whether the distributions of two populations are the same, (iii) whether two events might be independent, and (iv) whether there is a different variability than expected within a population.

An important parameter in a chi-square distribution is the degrees of freedom df in a given problem. The random variable in the chi-square distribution is the sum of squares of df standard normal variables, which must be independent. The key characteristics of the chi-square distribution also depend directly on the degrees of freedom.

The chi-square distribution curve is skewed to the right, and its shape depends on the degrees of freedom df. For df > 90, the curve approximates the normal distribution. Test statistics based on the chi-square distribution are always greater than or equal to zero.

Formula Review

$\chi^2 = (Z_1)^2 + (Z_2)^2 + \ldots (Z_{\text{df}})^2$ chi-square distribution random variable

$\mu_{\chi^2} = df$ chi-square distribution population mean

$\sigma_{\chi^2} = \sqrt{2\left(df\right)}$ Chi-Square distribution population standard deviation

Exercise:

Problem:

If the number of degrees of freedom for a chi-square distribution is 25, what is the population mean and standard deviation?

Solution:

mean = 25 and standard deviation = 7.0711

Exercise:

Problem:

If df > 90, the distribution is _______________. If df = 15, the distribution is ___________________.

Exercise:

Problem:

When does the chi-square curve approximate a normal distribution?

Solution:

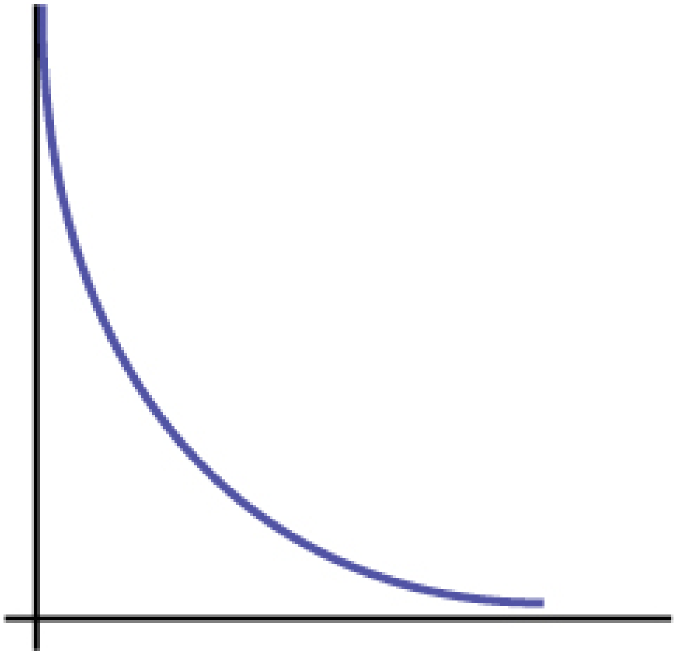
when the number of degrees of freedom is greater than 90

Exercise:

Problem: Where is μ located on a chi-square curve?

Exercise:

Problem: Is it more likely the df is 90, 20, or two in the graph?



Solution:

$df = 2$

Homework

Decide whether the following statements are true or false.

Exercise:

Problem:

As the number of degrees of freedom increases, the graph of the chi-square distribution looks more and more symmetrical.

Solution:

true

Exercise:

Problem:

The standard deviation of the chi-square distribution is twice the mean.

Exercise:

Problem:

The mean and the median of the chi-square distribution are the same if $df = 24$.

Solution:

false

Goodness-of-Fit Test

In this type of hypothesis test, you determine whether the data **"fit"** a particular distribution or not. For example, you may suspect your unknown data fit a binomial distribution. You use a chi-square test (meaning the distribution for the hypothesis test is chi-square) to determine if there is a fit or not. **The null and the alternative hypotheses for this test may be written in sentences or may be stated as equations or inequalities.**

The test statistic for a goodness-of-fit test is:
Equation:

$$\sum_k \frac{(O-E)^2}{E}$$

where:

- O = **observed values** (data)
- E = **expected values** (from theory)
- k = the number of different data cells or categories

The observed values are the data values and the expected values are the values you would expect to get if the null hypothesis were true. There are n terms of the form $\frac{(O-E)^2}{E}$.

The number of degrees of freedom is df = (number of categories – 1).

The **goodness-of-fit test** is always right-tailed. If the observed values and the corresponding expected values are not close to each other, then the test statistic can get very large and will be way out in the right tail of the chi-square curve.

Note:
Note
The expected value for each cell needs to be at least five in order for you to use this test.

Example:
Absenteeism of college students from math classes is a major concern to math instructors because missing class appears to increase the drop rate. Suppose that a study was done to determine if the actual student absenteeism rate follows faculty perception. The faculty expected that a group of 100 students would miss class according to the following table.

Number of absences per term	Expected number of students
0–2	50
3–5	30

Number of absences per term	Expected number of students
6–8	12
9–11	6
12+	2

A random survey across all mathematics courses was then done to determine the actual number **(observed)** of absences in a course. The chart in the table below displays the results of that survey.

Number of absences per term	Actual number of students
0–2	35
3–5	40
6–8	20
9–11	1
12+	4

Determine the null and alternative hypotheses needed to conduct a goodness-of-fit test.

H_0 : Student absenteeism **fits** faculty perception.

The alternative hypothesis is the opposite of the null hypothesis.

H_a : Student absenteeism **does not fit** faculty perception.

Exercise:

Problem: a. Can you use the information as it appears in the charts to conduct the goodness-of-fit test?

Solution:

a. **No.** Notice that the expected number of absences for the "12+" entry is less than five (it is two). Combine that group with the "9–11" group to create new tables where the number of students for each entry are at least five. The new results are in [link] and [link].

Number of absences per term	Expected number of students

Number of absences per term	Expected number of students
0–2	50
3–5	30
6–8	12
9+	8

Number of absences per term	Actual number of students
0–2	35
3–5	40
6–8	20
9+	5

Exercise:

Problem: b. What is the number of degrees of freedom (df)?

Solution:

b. There are four "cells" or categories in each of the new tables.

df = number of cells – 1 = 4 – 1 = 3

Note:
Try It
Exercise:

Problem:

A factory manager needs to understand how many products are defective versus how many are produced. The number of expected defects is listed in the following table.

Number produced	Number defective

Number produced	Number defective
0–100	5
101–200	6
201–300	7
301–400	8
401–500	10

A random sample was taken to determine the actual number of defects. The following table shows the results of the survey.

Number produced	Number defective
0–100	5
101–200	7
201–300	8
301–400	9
401–500	11

State the null and alternative hypotheses needed to conduct a goodness-of-fit test, and state the degrees of freedom.

Solution:

H_0 : The number of defaults fits expectations.

H_a : The number of defaults does not fit expectations.

$df = 4$

Example:
Exercise:

Problem:

Employers want to know which days of the week employees are absent in a five-day work week. Most employers would like to believe that employees are absent equally during the week. Suppose a random sample of 60 managers were asked on which day of the week they had the highest number of employee absences. The results were distributed as in the following table. For the population of employees, do the days for the highest number of absences occur with equal frequencies during a five-day work week? Test at a 5% significance level.

	Monday	**Tuesday**	**Wednesday**	**Thursday**	**Friday**
Number of Absences	15	12	9	9	15

Day of the Week Employees were Most Absent

Solution:

The null and alternative hypotheses are:

- H_0 : The absent days occur with equal frequencies, that is, they fit a uniform distribution.
- H_a : The absent days occur with unequal frequencies, that is, they do not fit a uniform distribution.

If the absent days occur with equal frequencies, then, out of 60 absent days (the total in the sample: 15 + 12 + 9 + 9 + 15 = 60), there would be 12 absences on Monday, 12 on Tuesday, 12 on Wednesday, 12 on Thursday, and 12 on Friday. These numbers are the **expected** (E) values. The values in the table are the **observed** (O) values or data.

This time, calculate the χ^2 test statistic by hand. Make a chart with the following headings and fill in the columns:

- Expected (E) values (12, 12, 12, 12, 12)
- Observed (O) values (15, 12, 9, 9, 15)
- $(O – E)$
- $(O – E)^2$
- $\frac{(O-E)^2}{E}$

Now add (sum) the last column. The sum is three. This is the χ^2 test statistic.

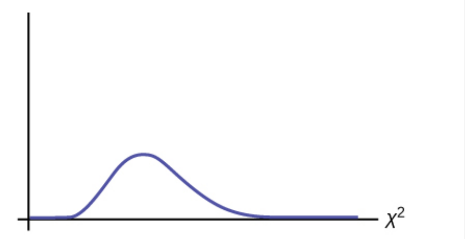
To find the p-value, calculate $P(\chi^2 > 3)$. This test is right-tailed. (Use a computer or calculator to find the p-value. You should get p-value = 0.5578.)

The degrees of freedom are the number of cells – 1 = 5 – 1 = 4

Note:

Press `2nd DISTR`. Arrow down to `χ²cdf`. Press `ENTER`. Enter `(3,1E99,4)`. Rounded to four decimal places, you should see 0.5578, which is the p-value.

Next, complete a graph like the following, one with the proper labeling and shading. (You should shade the right tail.)



The decision is not to reject the null hypothesis.

Conclusion: At a 5% level of significance, from the sample data, there is not sufficient evidence to conclude that the absent days do not occur with equal frequencies.

Note:
TI-83+ and some TI-84 calculators do not have a special program for the test statistic for the goodness-of-fit test. The next example [link] has the calculator instructions. The newer TI-84 calculators have in `STAT TESTS` the test `Chi2 GOF`. To run the test, put the observed values (the data) into a first list and the expected values (the values you expect if the null hypothesis is true) into a second list. Press `STAT TESTS` and `Chi2 GOF`. Enter the list names for the Observed list and the Expected list. Enter the degrees of freedom and press `calculate` or `draw`. Make sure you clear any lists before you start. **To Clear Lists in the calculators:** Go into `STAT EDIT` and arrow up to the list name area of the particular list. Press `CLEAR` and then arrow down. The list will be cleared. Alternatively, you can press `STAT` and press 4 (for `ClrList`). Enter the list name and press `ENTER`.

Note:
Try It
Exercise:

Problem:

Teachers want to know which night each week their students are doing most of their homework. Most teachers think that students do homework equally throughout the week. Suppose a random sample of 56 students were asked on which night of the week they did the most homework. The results were distributed as shown in the following table.

	Sunday	**Monday**	**Tuesday**	**Wednesday**	**Thursday**	**Friday**	**Saturda**
Number of Students	11	8	10	7	10	5	5

From the population of students, do the nights for the highest number of students doing the majority of their homework occur with equal frequencies during a week? What type of hypothesis test should you use?

Solution:

$df = 6$

p-value = 0.6093

We decline to reject the null hypothesis. There is not enough evidence to support that students do not do the majority of their homework equally throughout the week.

Example:

One study indicates that the number of televisions that American families have is distributed (this is the **given** distribution for the American population) as shown in the following table.

Number of Televisions	**Percent**
0	10
1	16
2	55
3	11
4+	8

The table contains expected (E) percents.

A random sample of 600 families in the far western United States resulted in the data in the following table.

Number of Televisions	**Frequency**
	Total = 600

Number of Televisions	**Frequency**
0	66
1	119
2	340
3	60
4+	15
	Total = 600

The table contains observed (O) frequency values.

Exercise:

Problem:

At the 1% significance level, does it appear that the distribution "number of televisions" of far western United States families is different from the distribution for the American population as a whole?

Solution:

This problem asks you to test whether the far western United States families distribution fits the distribution of the American families. This test is always right-tailed.

The first table contains expected percentages. To get expected (E) frequencies, multiply the percentage by 600. The expected frequencies are shown in the following table.

Number of Televisions	**Percent**	**Expected Frequency**
0	10	(0.10)(600) = 60
1	16	(0.16)(600) = 96
2	55	(0.55)(600) = 330
3	11	(0.11)(600) = 66
over 3	8	(0.08)(600) = 48

Therefore, the expected frequencies are 60, 96, 330, 66, and 48. In the TI calculators, you can let the calculator do the math. For example, instead of 60, enter 0.10*600.

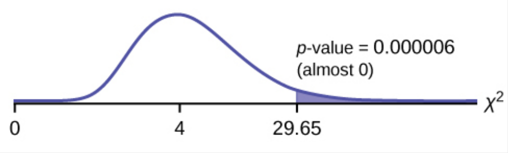
H_0 : The "number of televisions" distribution of far western United States families is the same as the "number of televisions" distribution of the American population.

H_a : The "number of televisions" distribution of far western United States families is different from the "number of televisions" distribution of the American population.

Distribution for the test: χ^2_4 where df = (the number of cells) – 1 = 5 – 1 = 4.

Calculate the test statistic: $\chi^2 = 29.65$

Graph:



Probability statement: p-value = $P(\chi^2 > 29.65) = 0.000006$

Compare α and the p-value:

- $\alpha = 0.01$
- p-value = 0.000006

So, the p-value < α.

Make a decision: Since the p-value < α, reject H_0 .

This means you reject the belief that the distribution for the far western states is the same as that of the American population as a whole.

Conclusion: At the 1% significance level, from the data, there is sufficient evidence to conclude that the "number of televisions" distribution for the far western United States is different from the "number of televisions" distribution for the American population as a whole.

Note:
Press `STAT` and `ENTER`. Make sure to clear lists `L1`, `L2`, and `L3` if they have data in them (see the note at the end of [link]). Into `L1`, put the observed frequencies `66`, `119`, `349`, `60`, `15`. Into `L2`, put the expected frequencies `.10*600, .16*600`, `.55*600`, `.11*600`, `.08*600`. Arrow over to list `L3` and up to the name area `"L3"`. Enter `(L1-L2)^2/L2` and `ENTER`. Press `2nd QUIT`. Press `2nd LIST` and arrow over to `MATH`. Press `5`. You should see `"sum" (Enter L3)`. Rounded to 2 decimal places, you should see `29.65`. Press `2nd DISTR`. Press `7` or Arrow down to `7:χ2cdf` and press `ENTER`. Enter `(29.65,1E99,4)`. Rounded to four places, you should see `5.77E-6 = .000006` (rounded to six decimal places), which is the p-value.
The newer TI-84 calculators have in `STAT TESTS` the test `Chi2 GOF`. To run the test, put the observed values (the data) into a first list and the expected values (the values you expect if the null hypothesis is true) into a second list. Press `STAT TESTS` and `Chi2 GOF`. Enter the list names for the Observed list and the Expected list. Enter the degrees of freedom and press `calculate` or `draw`. Make sure you clear any lists before you start.

Note:
Try It
Exercise:

Problem:

The expected percentage of the number of pets students have in their homes is distributed (this is the given distribution for the student population of the United States) as in the following table.

Number of Pets	**Percent**
0	18
1	25
2	30
3	18
4+	9

A random sample of 1,000 students from the Eastern United States resulted in the data in the following table.

Number of Pets	**Frequency**
0	210
1	240
2	320
3	140
4+	90

At the 1% significance level, does it appear that the distribution "number of pets" of students in the Eastern United States is different from the distribution for the United States student population as a whole? What is the p-value?

Solution:

p-value = 0.0036

We reject the null hypothesis that the distributions are the same. There is sufficient evidence to conclude that the distribution “number of pets” of students in the Eastern United States is different from the distribution for the United States student population as a whole.

Example:
Exercise:

Problem:

Suppose you flip two coins 100 times. The results are 20 *HH*, 27 *HT*, 30 *TH*, and 23 *TT*. Are the coins fair? Test at a 5% significance level.

Solution:

This problem can be set up as a goodness-of-fit problem. The sample space for flipping two fair coins is {*HH*, *HT*, *TH*, *TT*}. Out of 100 flips, you would expect 25 *HH*, 25 *HT*, 25 *TH*, and 25 *TT*. This is the expected distribution. The question, "Are the coins fair?" is the same as saying, "Does the distribution of the coins (20 *HH*, 27 *HT*, 30 *TH*, 23 *TT*) fit the expected distribution?"

Random Variable: Let X = the number of heads in one flip of the two coins. X takes on the values 0, 1, 2. (There are 0, 1, or 2 heads in the flip of two coins.) Therefore, the **number of cells is three**. Since X = the number of heads, the observed frequencies are 20 (for two heads), 57 (for one head), and 23 (for zero heads or both tails). The expected frequencies are 25 (for two heads), 50 (for one head), and 25 (for zero heads or both tails). This test is right-tailed.

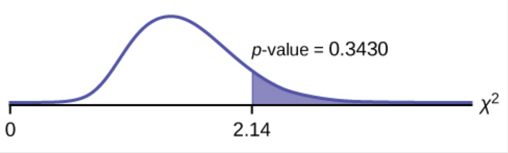
H_0 : The coins are fair.

H_a : The coins are not fair.

Distribution for the test: χ^2_2 where df = 3 – 1 = 2.

Calculate the test statistic: χ^2 = 2.14

Graph:



Probability statement: p-value = $P(\chi^2 > 2.14)$ = 0.3430

Compare α and the p-value:

- α = 0.05
- p-value = 0.3430

The p-value > α.

Make a decision: Since the p-value > α, do not reject H_0 .

Conclusion: There is insufficient evidence to conclude that the coins are not fair.

Note:
Press `STAT` and `ENTER`. Make sure you clear lists `L1`, `L2`, and `L3` if they have data in them. Into `L1`, put the observed frequencies `20`, `57`, `23`. Into `L2`, put the expected frequencies `25`, `50`, `25`. Arrow over to list `L3` and up to the name area `"L3"`. Enter `(L1-L2)^2/L2` and `ENTER`. Press `2nd QUIT`. Press `2nd LIST` and arrow over to `MATH`. Press `5`. You should see `"sum"`.`Enter L3`. Rounded to two decimal places, you should see `2.14`. Press `2nd DISTR`. Arrow down to `7:χ2cdf` (or press `7`). Press `ENTER`. Enter `2.14,1E99,2)`. Rounded to four places, you should see `.3430`, which is the p-value.

The newer TI-84 calculators have in `STAT TESTS` the test `Chi2 GOF`. To run the test, put the observed values (the data) into a first list and the expected values (the values you expect if the null hypothesis is true) into a second list. Press `STAT TESTS` and `Chi2 GOF`. Enter the list names for the Observed list and the Expected list. Enter the degrees of freedom and press `calculate` or `draw`. Make sure you clear any lists before you start.

Note:
Try It
Exercise:

Problem:

Students in a social studies class hypothesize that the literacy rates across the world for every region are 82%. The following table shows the actual literacy rates across the world broken down by region. What are the test statistic and the degrees of freedom?

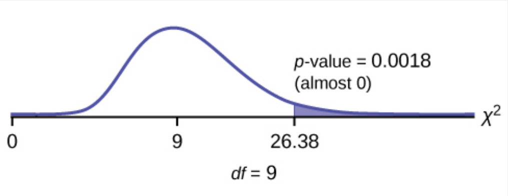
MDG Region	Adult Literacy Rate (%)
Developed Regions	99.0
Commonwealth of Independent States	99.5
Northern Africa	67.3
Sub-Saharan Africa	62.5
Latin America and the Caribbean	91.0

MDG Region	Adult Literacy Rate (%)
Eastern Asia	93.8
Southern Asia	61.9
South-Eastern Asia	91.9
Western Asia	84.5
Oceania	66.4

Solution:

degrees of freedom = 9

χ^2 test statistic = 26.38



Press `STAT` and `ENTER`. Make sure you clear lists `L1, L2,` and `L3` if they have data in them. Into L1, put the observed frequencies `99, 99.5, 67.3, 62.5, 91, 93.8, 61.9, 91.9, 84.5, 66.4`. Into `L2`, put the expected frequencies `82, 82, 82, 82, 82, 82, 82, 82, 82, 82`. Arrow over to list `L3` and up to the name area `"L3"`. Enter `(L1-L2)^2/L2` and `ENTER`. Press `2nd QUIT`. Press `2nd LIST` and arrow over to `MATH`. Press `5`. You should see `"sum"`. `Enter L3`. Rounded to two decimal places, you should see `26.38`. Press `2nd DISTR`. Arrow down to `7:χ2cdf` (or press `7`). Press `ENTER`. Enter `26.38,1E99,9)`. Rounded to four places, you should see `.0018`, which is the *p*-value.

The newer TI-84 calculators have in `STAT TESTS` the test `Chi2 GOF`. To run the test, put the observed values (the data) into a first list and the expected values (the values you expect if the null hypothesis is true) into a second list. Press `STAT TESTS` and `Chi2 GOF`. Enter the list names for the Observed list and the Expected list. Enter the degrees of freedom and press `calculate` or `draw`. Make sure you clear any lists before you start.

References

Data from the U.S. Census Bureau

Data from the College Board. Available online at http://www.collegeboard.com.

Data from the U.S. Census Bureau, Current Population Reports.

Ma, Y., E.R. Bertone, E.J. Stanek III, G.W. Reed, J.R. Hebert, N.L. Cohen, P.A. Merriam, I.S. Ockene, “Association between Eating Patterns and Obesity in a Free-living US Adult Population.” *American Journal of Epidemiology* volume 158, no. 1, pages 85-92.

Ogden, Cynthia L., Margaret D. Carroll, Brian K. Kit, Katherine M. Flegal, “Prevalence of Obesity in the United States, 2009–2010.” NCHS Data Brief no. 82, January 2012. Available online at http://www.cdc.gov/nchs/data/databriefs/db82.pdf (accessed May 24, 2013).

Stevens, Barbara J., “Multi-family and Commercial Solid Waste and Recycling Survey.” Arlington Count, VA. Available online at http://www.arlingtonva.us/departments/EnvironmentalServices/SW/file84429.pdf (accessed May 24,2013).

Section Review

To assess whether a data set fits a specific distribution, you can apply the goodness-of-fit hypothesis test that uses the chi-square distribution. The null hypothesis for this test states that the data come from the assumed distribution. The test compares observed values against the values you would expect to have if your data followed the assumed distribution. The test is always right-tailed. Each observation or cell category must have an expected value of at least five.

Formula Review

$\sum_{k} \frac{(O-E)^2}{E}$ goodness-of-fit test statistic where:

O: observed values
E: expected values

k: number of different data cells or categories

$df = k - 1$ degrees of freedom

Determine the appropriate test to be used in the next three exercises.

Exercise:

Problem:

An archeologist is calculating the distribution of the frequency of the number of artifacts she finds in a dig site. Based on previous digs, the archeologist creates an expected distribution broken down by grid sections in the dig site. Once the site has been fully excavated, she compares the actual number of artifacts found in each grid section to see if her expectation was accurate.

Exercise:

Problem:

An economist is deriving a model to predict outcomes on the stock market. He creates a list of expected points on the stock market index for the next two weeks. At the close of each day’s trading, he records the actual points on the index. He wants to see how well his model matched what actually happened.

Solution:

a goodness-of-fit test

Exercise:

Problem:

A personal trainer is putting together a weight-lifting program for her clients. For a 90-day program, she expects each client to lift a specific maximum weight each week. As she goes along, she records the actual maximum weights her clients lifted. She wants to know how well her expectations met with what was observed.

Use the following information to answer the next five exercises: A teacher predicts that the distribution of grades on the final exam will be and they are recorded in the following table.

Grade	**Proportion**
A	0.25
B	0.30
C	0.35
D	0.10

The actual distribution for a class of 60 is recorded in the following table.

Grade	**Frequency**
A	21
B	21
C	15
D	3

Exercise:

Problem: $df =$ ______

Solution:

3

Exercise:

Problem: State the null and alternative hypotheses.

Exercise:

Problem: χ^2 test statistic = ______

Solution:

6.11

Exercise:

Problem: p-value = ______

Exercise:

Problem: At the 5% significance level, what can you conclude?

Solution:

We decline to reject the null hypothesis. There is not enough evidence to suggest that the observed test scores are significantly different from the expected test scores.

Use the following information to answer the next nine exercises: The following data are real. The cumulative number of AIDS cases reported for Santa Clara County is broken down by ethnicity as in the following table.

Ethnicity	**Number of Cases**
White	2,229
Hispanic	1,157
Black/African-American	457
Asian, Pacific Islander	232
	Total = 4,075

The percentage of each ethnic group in Santa Clara County is shown in the table below.

Ethnicity	**Percentage of total county population**	**Number expected (round to two decimal places)**
White	42.9%	1748.18
Hispanic	26.7%	
Black/African-American	2.6%	
Asian, Pacific Islander	27.8%	
	Total = 100%	

Exercise:

Problem:

If the ethnicities of AIDS victims followed the ethnicities of the total county population, fill in the expected number of cases per ethnic group.
Perform a goodness-of-fit test to determine whether the occurrence of AIDS cases follows the ethnicities of the general population of Santa Clara County.

Exercise:

Problem: H_0 : _______

Solution:

H_0 : the distribution of AIDS cases follows the ethnicities of the general population of Santa Clara County.

Exercise:

Problem: H_a : _______

Exercise:

Problem: Is this a right-tailed, left-tailed, or two-tailed test?

Solution:

right-tailed

Exercise:

Problem: degrees of freedom = _______

Exercise:

Problem: χ^2 test statistic = _______

Solution:

88,621

Exercise:

Problem: p-value = _______

Exercise:

Problem:

Graph the situation. Label and scale the horizontal axis. Mark the mean and test statistic. Shade in the region corresponding to the p-value.

Let α = 0.05

Decision: _________________

Reason for the Decision: _________________

Conclusion (write out in complete sentences): _________________

Solution:

Graph: Check student’s solution.

Decision: Reject the null hypothesis.

Reason for the Decision: The p-value < alpha.

Conclusion (write out in complete sentences): The make-up of AIDS cases does not fit the ethnicities of the general population of Santa Clara County.

Exercise:

Problem:

Does it appear that the pattern of AIDS cases in Santa Clara County corresponds to the distribution of ethnic groups in this county? Why or why not?

Homework

For each problem, use a solution sheet to solve the hypothesis test problem. Go to [link] for the chi-square solution sheet. Round expected frequency to two decimal places.

Exercise:

Problem:

A six-sided die is rolled 120 times. Fill in the expected frequency column. Then, conduct a hypothesis test to determine if the die is fair. The data in the following table are the result of the 120 rolls.

Face Value	Frequency	Expected Frequency
1	15	
2	29	
3	16	
4	15	
5	30	
6	15	

Exercise:

Problem:

The marital status distribution of the U.S. male population, ages 15 and older, is as shown in the following table.

Marital Status	**Percent**	**Expected Frequency**
never married	31.3	
married	56.1	
widowed	2.5	
divorced/separated	10.1	

Suppose that a random sample of 400 U.S. young adult males, 18 to 24 years old, yielded the following frequency distribution. We are interested in whether this age group of males fits the distribution of the U.S. adult population. Calculate the frequency one would expect when surveying 400 people. Fill in [link], rounding each expected frequency to two decimal places.

Marital Status	**Frequency**
never married	140
married	238
widowed	2
divorced/separated	20

Solution:

Marital Status	**Percent**	**Expected Frequency**
never married	31.3	125.2
married	56.1	224.4
widowed	2.5	10

Marital Status	Percent	Expected Frequency
divorced/separated	10.1	40.4

a. The data fits the distribution.
b. The data does not fit the distribution.
c. 3
d. chi-square distribution with df = 3
e. 19.27
f. 0.0002
g. Check student's solution.

h.
  i. Alpha = 0.05
  ii. Decision: Reject null
  iii. Reason for decision: The p-value < alpha.
  iv. Conclusion: There is sufficient evidence to conclude that the data does not fit the distribution.

Use the following information to answer the next two exercises: The columns in the following table contain the Race/Ethnicity of U.S. Public Schools for a recent year, the percentages for the Advanced Placement Examinee Population for that class, and the Overall Student Population. Suppose the right column contains the result of a survey of 1,000 local students from that year who took an AP Exam.

Race/Ethnicity	AP Examinee Population	Overall Student Population	Survey Frequency
Asian, Asian American, or Pacific Islander	10.2%	5.4%	113
Black or African-American	8.2%	14.5%	94
Hispanic or Latino	15.5%	15.9%	136
American Indian or Alaska Native	0.6%	1.2%	10
White	59.4%	61.6%	604
Not reported/other	6.1%	1.4%	43

Exercise:

Problem:

Perform a goodness-of-fit test to determine whether the local results follow the distribution of the U.S. overall student population based on ethnicity.

Exercise:

Problem:

Perform a goodness-of-fit test to determine whether the local results follow the distribution of U.S. AP examinee population, based on ethnicity. Run the test at both the 5% and 1% level of significance.

Solution:

a. H_0 : The local results follow the distribution of the U.S. AP examinee population
b. H_a : The local results do not follow the distribution of the U.S. AP examinee population
c. df = 5
d. chi-square distribution with df = 5
e. chi-square test statistic = 13.4
f. p-value = 0.0199
g. Check student's solution.

h.
  i. Alpha = 0.05
  ii. Decision: Reject null when alpha is 0.05.
  iii. Reason for Decision: The p-value < alpha.
  iv. Conclusion: There is sufficient evidence to conclude that local data do not fit the AP examinee distribution.
  v. Alpha = 0.01
  vi. Decision: Do not reject null when alpha = 0.01
  vii. Reason for Decision: The p-value > alpha.
  viii. Conclusion: There is insufficient evidence to conclude that local data do not follow the distribution of the U.S. AP examinee distribution.

Exercise:

Problem:

The City of South Lake Tahoe, CA, has an Asian population of 1,419 people, out of a total population of 23,609. Suppose that a survey of 1,419 self-reported Asians in the Manhattan, NY, area yielded the data in the following table. Conduct a goodness-of-fit test to determine if the self-reported sub-groups of Asians in the Manhattan area fit that of the Lake Tahoe area.

Race	**Lake Tahoe Frequency**	**Manhattan Frequency**
Asian Indian	131	174
Chinese	118	557
Filipino	1,045	518
Japanese	80	54
Korean	12	29
Vietnamese	9	21
Other	24	66

Use the following information to answer the next two exercises: UCLA conducted a survey of more than 263,000 college freshmen from 385 colleges in fall 2005. The results of students' expected majors by gender were reported in *The Chronicle of Higher Education (2/2/2006)*. Suppose a survey of 5,000 graduating females and 5,000 graduating males was done as a follow-up last year to determine what their actual majors were. The results are shown in the tables for [link] and [link]. The second column in each table does not add to 100% because of rounding.

Exercise:

Problem:

Conduct a goodness-of-fit test to determine if the actual college majors of graduating females fit the distribution of their expected majors. Run the test at both the 5% and 1% level of significance.

Major	Women - Expected Major	Women - Actual Major
Arts & Humanities	14.0%	670
Biological Sciences	8.4%	410
Business	13.1%	685
Education	13.0%	650
Engineering	2.6%	145
Physical Sciences	2.6%	125
Professional	18.9%	975
Social Sciences	13.0%	605
Technical	0.4%	15
Other	5.8%	300
Undecided	8.0%	420

Solution:

a. H_0 : The actual college majors of graduating females fit the distribution of their expected majors
b. H_a : The actual college majors of graduating females do not fit the distribution of their expected majors
c. df = 10
d. chi-square distribution with df = 10
e. test statistic = 11.48
f. p-value = 0.3211
g. Check student’s solution.

h. i. Alpha = 0.05
  ii. Decision: Do not reject null when alpha = 0.05 and alpha = 0.01
  iii. Reason for decision: The p-value > alpha.

iv. Conclusion: There is insufficient evidence to conclude that the distribution of actual college majors of graduating females fits the distribution of their expected majors.

Exercise:

Problem:

Conduct a goodness-of-fit test to determine if the actual college majors of graduating males fit the distribution of their expected majors.

Major	**Men - Expected Major**	**Men - Actual Major**
Arts & Humanities	11.0%	600
Biological Sciences	6.7%	330
Business	22.7%	1130
Education	5.8%	305
Engineering	15.6%	800
Physical Sciences	3.6%	175
Professional	9.3%	460
Social Sciences	7.6%	370
Technical	1.8%	90
Other	8.2%	400
Undecided	6.6%	340

Read the statement and decide whether it is true or false.

Exercise:

Problem:

In a goodness-of-fit test, the expected values are the values we would expect if the null hypothesis were true.

Solution:

true

Exercise:

Problem:

In general, if the observed values and expected values of a goodness-of-fit test are not close together, then the test statistic can get very large and on a graph will be way out in the right tail.

Exercise:

Problem:

Use a goodness-of-fit test to determine if high school principals believe that students are absent equally during the week or not.

Solution:

true

Exercise:

Problem: The test to use to determine if a six-sided die is fair is a goodness-of-fit test.

Exercise:

Problem: In a goodness-of fit test, if the p-value is 0.0113, in general, do not reject the null hypothesis.

Solution:

false

Exercise:

Problem:

A sample of 212 commercial businesses was surveyed for recycling one commodity; a commodity here means any one type of recyclable material such as plastic or aluminum. The following table shows the business categories in the survey, the sample size of each category, and the number of businesses in each category that recycle one commodity. Based on the study, on average half of the businesses were expected to be recycling one commodity. As a result, the last column shows the expected number of businesses in each category that recycle one commodity. At the 5% significance level, perform a hypothesis test to determine if the observed number of businesses that recycle one commodity follows the uniform distribution of the expected values.

Business Type	Number in class	Observed Number that recycle one commodity	Expected number that recycle one commodity
Office	35	19	17.5
Retail/Wholesale	48	27	24
Food/Restaurants	53	35	26.5
Manufacturing/Medical	52	21	26
Hotel/Mixed	24	9	12

Exercise:

Problem:

The following table contains information from a survey among 499 participants classified according to their age groups. The second column shows the percentage of obese people per age class among the study participants. The last column comes from a different study at the national level that shows the corresponding percentages of obese people in the same age classes in the USA. Perform a hypothesis test at the 5% significance level to determine whether the survey participants are a representative sample of the USA obese population.

Age Class (Years)	Obese (Percentage)	Expected USA average (Percentage)
20–30	75.0	32.6
31–40	26.5	32.6
41–50	13.6	36.6
51–60	21.9	36.6
61–70	21.0	39.7

Solution:

a. H_0 : Surveyed obese fit the distribution of expected obese
b. H_a : Surveyed obese do not fit the distribution of expected obese
c. df = 4
d. chi-square distribution with df = 4
e. test statistic = 54.01
f. p-value = 0
g. Check student’s solution.

h.
   i. Alpha: 0.05
   ii. Decision: Reject the null hypothesis.
   iii. Reason for decision: The p-value < alpha.
   iv. Conclusion: At the 5% level of significance, from the data, there is sufficient evidence to conclude that the surveyed obese do not fit the distribution of expected obese.

Glossary

Goodness-of-fit Test
: a hypothesis test used to determine whether a set of data fit a particular distribution

Observed values
: data values collected from the sample

Expected values
: values you would expect if the null hypothesis were true

Test of Independence

A test of independence determines whether two factors (or events) are independent or not. You first encountered the term independence in Chapter 3.

Tests of independence involve using a **contingency table** of observed (data) values. Contingency tables were also covered in Chapter 3.

The test statistic for a **test of independence** is similar to that of a goodness-of-fit test:
Equation:

$$\chi^2 = \sum_{(i \cdot j)} \frac{(O - E)^2}{E}$$

where:

- O = observed values
- E = expected values
- i = the number of rows in the table
- j = the number of columns in the table

There are $i \cdot j$ terms of the form $\frac{(O-E)^2}{E}$.

Note:
Note
The expected value for each cell needs to be at least five in order for you to use this test.

The following illustrates an example of a test for independence.

Example:
Suppose A = a speeding violation in the last year and B = a cell phone user while driving. If A and B are independent then $P(A \text{ AND } B) = P(A)P(B)$. The event A AND B is the event that a driver received a speeding violation last year and also used a cell phone while driving. Suppose, in a study of drivers who received speeding violations in the last year, and who used cell phone while driving, that 755 people were surveyed. Out of the 755, 70 had a speeding violation and 685 did not; 305 used cell phones while driving and 450 did not. These observed values are shown in the following contingency table.

	Speeding violation in the last year	**No speeding violation in the last year**	**Total**
Cell phone user	25	280	305

	Speeding violation in the last year	**No speeding violation in the last year**	**Total**
Not a cell phone user	45	405	450
Total	70	685	755

Let y = expected number of drivers who used a cell phone while driving and received speeding violations.

If A and B are independent, then $P(A \text{ AND } B) = P(A)P(B)$. By substitution,

$$\frac{y}{755} = \left(\frac{70}{755}\right)\left(\frac{305}{755}\right)$$

Solve for y: $y = \frac{(70)(305)}{755} = 28.3$

About 28 people from the sample are expected to use cell phones while driving and to receive speeding violations. In a test of independence, we state the null and alternative hypotheses in words. The null hypothesis states that the events are **independent** and the alternative hypothesis states that they are **not independent (dependent)**. If we do a test of independence using this example, then the null hypothesis is:

H_0 : Being a cell phone user while driving and receiving a speeding violation are independent events.

If the null hypothesis were true, we would expect about 28 people to use cell phones while driving and to receive a speeding violation.

The test of independence is always right-tailed because of the calculation of the test statistic. If the expected and observed values are not close together, then the test statistic is very large and way out in the right tail of the chi-square curve, as it is in a goodness-of-fit.

The number of degrees of freedom for the test of independence is:

df = (number of columns - 1)(number of rows - 1)

The following formula calculates the **expected number** (E):

$$E = \frac{(\text{row total})(\text{column total})}{\text{total number surveyed}}$$

Note:
Try It
Exercise:

Problem:

A sample of 300 students is taken. Of the students surveyed, 50 were music students, while 250 were not. Ninety-seven were on the honor roll, while 203 were not. If we assume being a music student and being on the honor roll are independent events, what is the expected number of music students who are also on the honor roll? (It may help to make the contingency table.)

Solution:

About 16 students are expected to be music students and on the honor roll.

Example:

In a volunteer group, adults 21 and older volunteer from one to nine hours each week to spend time with a disabled senior citizen. The program recruits among community college students, four-year college students, and nonstudents. In the following table is a **sample** of the adult volunteers and the number of hours they volunteer per week.

Type of Volunteer	**1–3 Hours**	**4–6 Hours**	**7–9 Hours**	**Row Total**
Community College Students	111	96	48	255
Four-Year College Students	96	133	61	290
Nonstudents	91	150	53	294
Column Total	298	379	162	839

Number of Hours Worked Per Week by Volunteer Type (Observed)

Exercise:

Problem: Is the number of hours volunteered **independent** of the type of volunteer?

Solution:

The **observed table** and the question at the end of the problem, "Is the number of hours volunteered independent of the type of volunteer?" tell you this is a test of independence. The two factors are **number of hours volunteered** and **type of volunteer**. This test is always right-tailed.

H_0 : The number of hours volunteered is **independent** of the type of volunteer.

H_a : The number of hours volunteered is **dependent** on the type of volunteer.

The expected results are in the following table.

Type of Volunteer	**1-3 Hours**	**4-6 Hours**	**7-9 Hours**

Type of Volunteer	1-3 Hours	4-6 Hours	7-9 Hours
Community College Students	90.57	115.19	49.24
Four-Year College Students	103.00	131.00	56.00
Nonstudents	104.42	132.81	56.77

Number of Hours Worked Per Week by Volunteer Type (Expected)

For example, the calculation for the expected frequency for the top left cell is

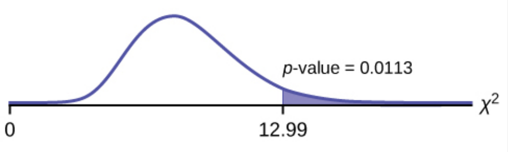
$$E = \frac{(\text{row total})(\text{column total})}{\text{total number surveyed}} = \frac{(255)\,(298)}{839} = 90.57$$

df = (3 columns – 1)(3 rows – 1) = (2)(2) = 4

Calculate the test statistic: χ^2 = 12.99 (calculator or computer)

Distribution for the test: χ^2_4

Graph:



Probability statement: p-value = $P(\chi^2 > 12.99)$ = 0.0113

Compare α and the p-value: Since no α is given, assume α = 0.05. The p-value = 0.0113. So the p-value < α.

Make a decision: Since the p-value < α, reject H_0 . This means that the events are not independent.

Conclusion: At a 5% level of significance, from the data, there is sufficient evidence to conclude that the number of hours volunteered and the type of volunteer are dependent on one another.

For the example in [link], if there had been another type of volunteer, teenagers, what would the degrees of freedom be?

Note:
Press the `MATRX` key and arrow over to `EDIT`. Press `1:[A]`. Press `3 ENTER 3 ENTER`. Enter the table values by row from [link]. Press `ENTER` after each. Press `2nd QUIT`. Press `STAT` and arrow over to `TESTS`. Arrow down to `C:χ2-TEST`. Press `ENTER`. You should see `Observed:[A] and`

`Expected:[B]`. Arrow down to `Calculate`. Press `ENTER`. The test statistic is 12.9909 and the *p*-value = 0.0113. Do the procedure a second time, but arrow down to `Draw` instead of `calculate`.

Note: The calculator automatically generates the expected values and stores them under matrix B. After running the test, you may look up these expected values by pressing the

`MATRX`

key, arrowing over to

`EDIT`

, and then entering

`2:[B]`

.

Note:
Try It
Exercise:

Problem:

The Bureau of Labor Statistics gathers data about employment in the United States. A sample is taken to calculate the number of U.S. citizens working in one of several industry sectors over time. The following table shows the results:

Industry Sector	**2000**	**2010**	**2020**	**Total**
Nonagriculture wage and salary	13,243	13,044	15,018	41,305
Goods-producing, excluding agriculture	2,457	1,771	1,950	6,178
Services-providing	10,786	11,273	13,068	35,127
Agriculture, forestry, fishing, and hunting	240	214	201	655
Nonagriculture self-employed and unpaid family worker	931	894	972	2,797
Secondary wage and salary jobs in agriculture and private household industries	14	11	11	36

Industry Sector	**2000**	**2010**	**2020**	**Total**
Secondary jobs as a self-employed or unpaid family worker	196	144	152	492
Total	27,867	27,351	31,372	86,590

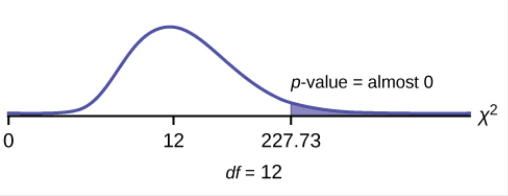
We want to know if the change in the number of jobs is independent of the change in years. State the null and alternative hypotheses and the degrees of freedom. Then run a test of independence.

Solution:

H_0 : The number of jobs is independent of the year.

H_a : The number of jobs is dependent on the year.

$df = 12$



Press the `MATRX` key and arrow over to `EDIT`. Press `1:[A]`. Press `3 ENTER 3 ENTER`. Enter the table values by row. Press `ENTER` after each. Press `2nd QUIT`. Press `STAT` and arrow over to `TESTS`. Arrow down to `C:χ2-TEST`. Press `ENTER`. You should see `Observed:[A] and Expected:[B]`. Arrow down to `Calculate`. Press `ENTER`. The test statistic is 227.73 and the p−value = 5.90E - 42 = 0. Do the procedure a second time but arrow down to `Draw` instead of `calculate`.

Conclusion: There is sufficient evidence to conclude that the number of jobs is dependent on the year.

Example:
De Anza College is interested in the relationship between anxiety level and the need to succeed in school. A random sample of 400 students took a test that measured anxiety level and need to succeed in school. The following table shows the results. De Anza College wants to know if anxiety level and the need to succeed in school are independent events.

Need to Succeed in School	**High Anxiety**	**Med-high Anxiety**	**Medium Anxiety**	**Med-low Anxiety**	**Low Anxiety**	**Row Total**
High Need	35	42	53	15	10	155
Medium Need	18	48	63	33	31	193
Low Need	4	5	11	15	17	52

Need to Succeed in School	**High Anxiety**	**Med-high Anxiety**	**Medium Anxiety**	**Med-low Anxiety**	**Low Anxiety**	**Row Total**
Column Total	57	95	127	63	58	400

Need to Succeed in School vs. Anxiety Level

Exercise:

Problem: a. How many high anxiety level students are expected to have a high need to succeed in school?

Solution:

a. The column total for a high anxiety level is 57. The row total for high need to succeed in school is 155. The sample size or total surveyed is 400.

$$E=\frac{\text{(row total)(column total)}}{\text{total surveyed}}=\frac{155\cdot 57}{400}=22.09$$

The expected number of students who have a high anxiety level and a high need to succeed in school is about 22.

Exercise:

Problem:

b. If the two variables are independent, how many students do you expect to have a low need to succeed in school and a med-low level of anxiety?

Solution:

b. The column total for a med-low anxiety level is 63. The row total for a low need to succeed in school is 52. The sample size or total surveyed is 400.

Exercise:

Problem: c. $E=\frac{\text{(row total)(column total)}}{\text{total surveyed}}=$ ________

Solution:

c. $E=\frac{\text{(row total)(column total)}}{\text{total surveyed}}=8.19$

Exercise:

Problem:

d. The expected number of students who have a med-low anxiety level and a low need to succeed in school is about ________.

Solution:

d. 8

Note:
Try It
Exercise:

Problem:

Refer back to the information in [link]. How many service providing jobs are there expected to be in 2020? How many nonagriculture wage and salary jobs are there expected to be in 2020?

Solution:

12,727, 14,965

References

DiCamilo, Mark, Mervin Field, “Most Californians See a Direct Linkage between Obesity and Sugary Sodas. Two in Three Voters Support Taxing Sugar-Sweetened Beverages If Proceeds are Tied to Improving School Nutrition and Physical Activity Programs.” The Field Poll, released Feb. 14, 2013. Available online at http://field.com/fieldpollonline/subscribers/Rls2436.pdf (accessed May 24, 2013).

Harris Interactive, “Favorite Flavor of Ice Cream.” Available online at http://www.statisticbrain.com/favorite-flavor-of-ice-cream (accessed May 24, 2013)

“Youngest Online Entrepreneurs List.” Available online at http://www.statisticbrain.com/youngest-online-entrepreneur-list (accessed May 24, 2013).

Section Review

To assess whether two factors are independent or not, you can apply the test of independence that uses the chi-square distribution. The null hypothesis for this test states that the two factors are independent. The test compares observed values to expected values. The test is right-tailed. Each observation or cell category must have an expected value of at least 5.

Formula Review

Test of Independence

- The number of degrees of freedom is equal to (number of columns - 1)(number of rows - 1).
- The test statistic is $\chi^2 = \sum\limits_{(i \cdot j)} \frac{(O-E)^2}{E}$ where O = observed values, E = expected values, i = the number of rows in the table, and j = the number of columns in the table.
- If the null hypothesis is true, the expected number $E = \frac{\text{(row total)(column total)}}{\text{total surveyed}}$.

Determine the appropriate test to be used in the next three exercises.
Exercise:

Problem:

A pharmaceutical company is interested in the relationship between age and presentation of symptoms for a common viral infection. A random sample is taken of 500 people with the infection across different age groups.

Solution:

a test of independence

Exercise:

Problem:

The owner of a baseball team is interested in the relationship between player salaries and team winning percentage. He takes a random sample of 100 players from different organizations.

Exercise:

Problem:

A marathon runner is interested in the relationship between the brand of shoes runners wear and their run times. She takes a random sample of 50 runners and records their run times as well as the brand of shoes they were wearing.

Solution:

a test of independence

Use the following information to answer the next seven exercises: Transit Railroads is interested in the relationship between travel distance and the ticket class purchased. A random sample of 200 passengers is taken. The following table shows the results. The railroad wants to know if a passenger's choice in ticket class is independent of the distance they must travel.

Traveling Distance	**Third class**	**Second class**	**First class**	**Total**
1–100 miles	21	14	6	41
101–200 miles	18	16	8	42
201–300 miles	16	17	15	48
301–400 miles	12	14	21	47
401–500 miles	6	6	10	22
Total	73	67	60	200

Exercise:

Problem: State the hypotheses.
H_0 : ________
H_a : ________

Exercise:

Problem: degrees of freedom = ________

Solution:

8

Exercise:

Problem:

How many passengers are expected to travel between 201 and 300 miles and purchase second-class tickets?

Exercise:

Problem:

How many passengers are expected to travel between 401 and 500 miles and purchase first-class tickets?

Solution:

6.6

Exercise:

Problem: What is the test statistic?

Exercise:

Problem: What is the p-value?

Solution:

0.0435

Exercise:

Problem: What can you conclude at the 5% level of significance?

Use the following information to answer the next eight exercises: An article in the New England Journal of Medicine, discussed a study on smokers in California and Hawaii. In one part of the report, the self-reported ethnicity and smoking levels per day were given. Of the people smoking at most ten cigarettes per day, there were 9,886 African Americans, 2,745 Native Hawaiians, 12,831 Latinos, 8,378 Japanese Americans and 7,650 whites. Of the people smoking 11 to 20 cigarettes per day, there were 6,514 African Americans, 3,062 Native Hawaiians, 4,932 Latinos, 10,680 Japanese Americans, and 9,877 whites. Of the people smoking 21 to 30 cigarettes per day, there were 1,671 African Americans, 1,419 Native Hawaiians, 1,406 Latinos, 4,715 Japanese Americans, and 6,062 whites. Of the people smoking at least 31 cigarettes per day, there were 759 African Americans, 788 Native Hawaiians, 800 Latinos, 2,305 Japanese Americans, and 3,970 whites.

Exercise:

Problem: Complete the table.

Smoking Level Per Day	**African American**	**Native Hawaiian**	**Latino**	**Japanese Americans**	**White**	**TOTALS**
1-10						

Smoking Level Per Day	**African American**	**Native Hawaiian**	**Latino**	**Japanese Americans**	**White**	**TOTALS**
11-20						
21-30						
31+						
TOTALS						

Smoking Levels by Ethnicity (Observed)

Solution:

Smoking Level Per Day	**African American**	**Native Hawaiian**	**Latino**	**Japanese Americans**	**White**	**Totals**
1-10	9,886	2,745	12,831	8,378	7,650	41,490
11-20	6,514	3,062	4,932	10,680	9,877	35,065
21-30	1,671	1,419	1,406	4,715	6,062	15,273
31+	759	788	800	2,305	3,970	8,622
Totals	18,830	8,014	19,969	26,078	27,559	10,0450

Exercise:

State the hypotheses.

H_0 : ________

Problem: H_a : ________

Exercise:

Problem: Make a table of expected values for [link]. Round to two decimal places.

Solution:

Smoking Level Per Day	**African American**	**Native Hawaiian**	**Latino**	**Japanese Americans**	**White**
1-10	7777.57	3310.11	8248.02	10771.29	11383.01
11-20	6573.16	2797.52	6970.76	9103.29	9620.27
21-30	2863.02	1218.49	3036.20	3965.05	4190.23
31+	1616.25	687.87	1714.01	2238.37	2365.49

Exercise:

Problem: degrees of freedom = ________

Exercise:

Problem: χ^2 test statistic = _______

Solution:

10,301.8

Exercise:

Problem: p-value = _______

Exercise:

Problem: Is this a right-tailed, left-tailed, or two-tailed test?

Solution:

right

Exercise:

Problem:

Graph the situation. Label and scale the horizontal axis. Mark the mean and test statistic. Shade in the region corresponding to the p-value.

State the decision and conclusion (in a complete sentence) for the following preconceived levels of α.

Exercise:

Problem: α = 0.05

a. Decision: ____________________
b. Reason for the decision: ____________________
c. Conclusion (write out in a complete sentence): ____________________

Solution:

a. Reject the null hypothesis.
b. The p-value < alpha.
c. There is sufficient evidence to conclude that smoking level is dependent on ethnic group.

Exercise:

Problem: $\alpha = 0.01$

a. Decision: ____________________
b. Reason for the decision: ____________________
c. Conclusion (write out in a complete sentence): ____________________

Homework

For each problem, use a solution sheet to solve the hypothesis test problem. Go to Appendix A for the Solution Sheets. Round expected frequency to two decimal places.

Exercise:

Problem:

A recent debate about where in the United States skiers believe the skiing is best prompted the following survey. Test to see if the best ski area is independent of the level of the skier.

U.S. Ski Area	Beginner	Intermediate	Advanced
Tahoe	20	30	40
Utah	10	30	60
Colorado	10	40	50

Exercise:

Problem:

Car manufacturers are interested in whether there is a relationship between the size of car an individual drives and the number of people in the driver's family (that is, whether car size and family size are independent). To test this, suppose that 800 car owners were randomly surveyed with the results in the following table. Conduct a test of independence.

Family Size	**Sub & Compact**	**Mid-size**	**Full-size**	**Van & Truck**
1	20	35	40	35
2	20	50	70	80
3–4	20	50	100	90
5+	20	30	70	70

Solution:

a. H_0 : Car size is independent of family size.
b. H_a : Car size is dependent on family size.
c. $df = 9$
d. chi-square distribution with $df = 9$
e. test statistic: $\chi^2 = 15.8284$
f. p-value = 0.0706
g. Check student's solution.

h.
  i. Alpha: 0.05
  ii. Decision: Do not reject the null hypothesis.
  iii. Reason for decision: The p-value > alpha.
  iv. Conclusion: At the 5% significance level, there is insufficient evidence to conclude that car size and family size are dependent.

Exercise:

Problem:

College students may be interested in whether or not their majors have any effect on starting salaries after graduation. Suppose that 300 recent graduates were surveyed as to their majors in college and their starting salaries after graduation. The following table shows the observed data. Conduct a test of independence.

Major	**< $50,000**	**$50,000 – $68,999**	**$69,000 +**
English	5	20	5
Engineering	10	30	60
Nursing	10	15	15
Business	10	20	30
Psychology	20	30	20

Exercise:

Problem:

Some travel agents claim that honeymoon hot spots vary according to age of the bride. Suppose that 280 recent brides were interviewed as to where they spent their honeymoons. The information is given in the following table. Conduct a test of independence.

Location	**20–29**	**30–39**	**40–49**	**50 and over**
Niagara Falls	15	25	25	20
Poconos	15	25	25	10
Europe	10	25	15	5
Virgin Islands	20	25	15	5

Solution:

a. H_0 : Honeymoon locations are independent of bride’s age.
b. H_a : Honeymoon locations are dependent on bride’s age.
c. df = 9
d. chi-square distribution with df = 9
e. test statistic: χ^2 = 15.7027
f. p-value = 0.0734
g. Check student’s solution.

h.
  i. Alpha: 0.05
  ii. Decision: Do not reject the null hypothesis.
  iii. Reason for decision: The p-value > alpha.
  iv. Conclusion: At the 5% significance level, there is insufficient evidence to conclude that honeymoon location is dependent on the bride's age.

Exercise:

Problem:

A manager of a sports club keeps information concerning the main sport in which members participate and their ages. To test whether there is a relationship between the age of a member and his or her choice of sport, 643 members of the sports club are randomly selected. Conduct a test of independence.

Sport	**18 - 25**	**26 - 30**	**31 - 40**	**41 and over**
racquetball	42	58	30	46
tennis	58	76	38	65

Sport	**18 - 25**	**26 - 30**	**31 - 40**	**41 and over**
swimming	72	60	65	33

Exercise:

Problem:

A major food manufacturer is concerned that the sales for its skinny french fries have been decreasing. As a part of a feasibility study, the company conducts research into the types of fries sold across the country to determine if the type of fries sold is independent of the area of the country. The results of the study are shown in the following table. Conduct a test of independence.

Type of Fries	**Northeast**	**South**	**Central**	**West**
skinny fries	70	50	20	25
curly fries	100	60	15	30
steak fries	20	40	10	10

Solution:

a. H_0 : The types of fries sold are independent of the location.
b. H_a : The types of fries sold are dependent on the location.
c. df = 6
d. chi-square distribution with df = 6
e. test statistic: χ^2 = 18.8369
f. p-value = 0.0044
g. Check student's solution.

h.
   i. Alpha: 0.05
   ii. Decision: Reject the null hypothesis.
   iii. Reason for decision: The p-value < alpha.
   iv. Conclusion: At the 5% significance level, There is sufficient evidence that types of fries and location are dependent.

Exercise:

Problem:

According to Dan Lenard, an independent insurance agent in the Buffalo, N.Y. area, the following is a breakdown of the amount of life insurance purchased by males in the following age groups. He is interested in whether the age of the male and the amount of life insurance purchased are independent events. Conduct a test for independence.

Age of Males	None	< $200,000	$200,000–$400,000	$401,001–$1,000,000	$1,000,001+
20–29	40	15	40	0	5
30–39	35	5	20	20	10
40–49	20	0	30	0	30
50+	40	30	15	15	10

Exercise:

Problem:

Suppose that 600 thirty-year-olds were surveyed to determine whether or not there is a relationship between the level of education an individual has and salary. Conduct a test of independence.

Annual Salary	Not a high school graduate	High school graduate	College graduate	Masters or doctorate
< $30,000	15	25	10	5
$30,000–$40,000	20	40	70	30
$40,000–$50,000	10	20	40	55
$50,000–$60,000	5	10	20	60
$60,000+	0	5	10	150

Solution:

a. H_0 : Salary is independent of level of education.
b. H_a : Salary is dependent on level of education.
c. df = 12
d. chi-square distribution with df = 12
e. test statistic: χ^2 = 255.7704
f. p-value = 0
g. Check student’s solution.

h.
   i. Alpha: 0.05
   ii. Decision: Reject the null hypothesis.
   iii. Reason for decision: The p-value < alpha.
   iv. Conclusion: At the 5% significance level, there is sufficient evidence to conclude that salary and level of education are dependent.

Read the statement and decide whether it is true or false.

Exercise:

Problem: The number of degrees of freedom for a test of independence is equal to the sample size minus one.

Exercise:

Problem: The test for independence uses tables of observed and expected data values.

Solution:

true

Exercise:

Problem:

The test to use when determining if the college or university a student chooses to attend is related to his or her socioeconomic status is a test for independence.

Exercise:

Problem:

In a test of independence, the expected number is equal to the row total multiplied by the column total divided by the total surveyed.

Solution:

true

Exercise:

Problem:

An ice cream maker performs a nationwide survey about favorite flavors of ice cream in different geographic areas of the U.S. Based on the following table, do the numbers suggest that geographic location is independent of favorite ice cream flavors? Test at the 5% significance level.

U.S. region/Flavor	Strawberry	Chocolate	Vanilla	Rocky Road	Mint Chocolate Chip	Pistachio
West	12	21	22	19	15	8
Midwest	10	32	22	11	15	6
East	8	31	27	8	15	7
South	15	28	30	8	15	6
Column Total	45	112	101	46	60	27

Exercise:

Problem:

The following table provides a recent survey of the youngest online entrepreneurs whose net worth is estimated at one million dollars or more. Their ages range from 17 to 30. Each cell in the table illustrates the number of entrepreneurs who correspond to the specific age group and their net worth. Are the ages and net worth independent? Perform a test of independence at the 5% significance level.

Age Group\ Net Worth Value (in millions of US dollars)	**1–5**	**6–24**	**≥25**	**Row Total**
17–25	8	7	5	20
26–30	6	5	9	20
Column Total	14	12	14	40

Solution:

a. H_0 : Age is independent of the youngest online entrepreneurs' net worth.
b. H_a : Age is dependent on the net worth of the youngest online entrepreneurs.
c. df = 2
d. chi-square distribution with df = 2
e. test statistic: χ^2 = 1.76
f. p-value = 0.4144
g. Check student's solution.

h.
  i. Alpha: 0.05
  ii. Decision: Do not reject the null hypothesis.
  iii. Reason for decision: The p-value > alpha.
  iv. Conclusion: At the 5% significance level, there is insufficient evidence to conclude that age and net worth for the youngest online entrepreneurs are dependent.

Exercise:

Problem:

A 2013 poll in California surveyed people about taxing sugar-sweetened beverages. The results are presented in the following table, and are classified by ethnic group and response type. Are the poll responses independent of the participants' ethnic group? Conduct a test of independence at the 5% significance level.

Opinion/Ethnicity	**Asian-American**	**White/Non-Hispanic**	**African-American**	**Latino**	**Row Total**
Against tax	48	433	41	160	628

Opinion/Ethnicity	Asian-American	White/Non-Hispanic	African-American	Latino	Row Total
In Favor of tax	54	234	24	147	459
No opinion	16	43	16	19	84
Column Total	118	710	71	272	1171

Glossary

Contingency Table

a table that displays sample values for two different factors that may be dependent or contingent on one another; it facilitates determining conditional probabilities.

Linear Regression and Correlation: Introduction class="introduction"

Linear regression and correlation can help you determine if an auto mechanic’s salary is related to his work experience. (credit: Joshua Rothhaas)

Note:
Chapter Objectives
By the end of this chapter, the student should be able to:

- Discuss basic ideas of linear regression and correlation.
- Create and interpret a line of best fit.
- Use the equation of the line of best fit to make predictions.
- Calculate and interpret the correlation coefficient.
- Test the significance of the correlation coefficient. (optional)

Professionals often want to know how two or more numeric variables are related. For example, is there a relationship between the grade on the second math exam a student takes and the grade on the final exam? If there is a relationship, what is the relationship and how strong is it?

In another example, your income may be determined by your education, your profession, your years of experience, and your ability. The amount you pay a repair person for labor is often determined by an initial amount plus an hourly fee.

The type of data described in the examples is **bivariate** data — "bi" for two variables. In reality, statisticians use **multivariate** data, meaning many variables.

In this chapter, you will be studying the simplest form of regression, "linear regression" with one independent variable ( ). This involves data that fits a line in two dimensions. You will also study correlation which measures how strong the relationship is.

Linear Equations

The most basic type of association is a linear association. This type of relationship can be defined algebraically by the equations used, numerically with actual or predicted data values, or graphically from a plotted curve. (Lines are classified as straight curves.)

Linear regression for two variables is based on a **linear equation** with one independent variable. The equation has the form:

Equation:

$$y = a + bx$$

where a and b are constant numbers.

The variable x **is the independent variable, and** y **is the dependent variable.** Typically, you choose a value to substitute for the independent variable and then solve for the dependent variable.

Example:

The following examples are linear equations.

Equation:

$$y = 2 + 3x$$

Equation:

$$y = 1.2x - 0.01$$

Note:
Try It
Exercise:

Problem: Is the following an example of a linear equation?

$y = -0.125 - 3.5x$

Solution:

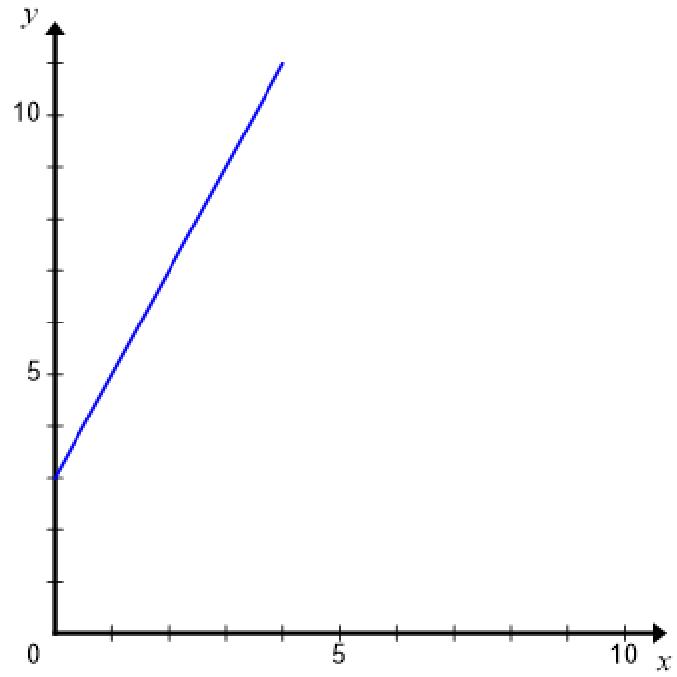
Yes. It is in the form $y = a + bx$.

The graph of a linear equation of the form $y = a + bx$ is a **straight line**. Any line that is not vertical can be described by this equation.

b is the slope of the line and a is the y-coordinate of the y-intercept.

Example:

Graph the equation $y = 2x + \ 3$.



Note that the line crosses the y-axis at (0, 3) and has a slope of 2, which means it rises 2 units for every 1 unit it moves to the right.

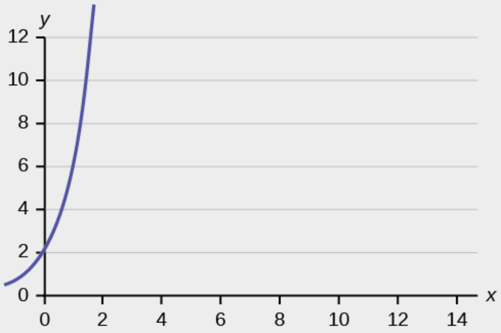
Note:

Try It

Exercise:

Problem:

Is the following the graph of a linear equation? Why or why not?



Solution:

No, the graph is not a straight line; therefore, it cannot be the graph of a linear equation.

Example:

Aaron's Word Processing Service (AWPS) does word processing. The rate for services is $32 per hour plus a $31.50 one-time charge. The total cost to a customer depends on the number of hours it takes to complete the job.

Exercise:

Problem:

Find the equation that expresses the **total cost** in terms of the **number of hours** required to complete the job.

Solution:

Let x = the number of hours it takes to get the job done.
Let y = the total cost to the customer.

The $31.50 is a fixed cost. If it takes x hours to complete the job, then (32)(x) is the cost of the word processing only. The total cost is: $y = 31.5 + 32x$

Note:
Try It
Exercise:

Problem:

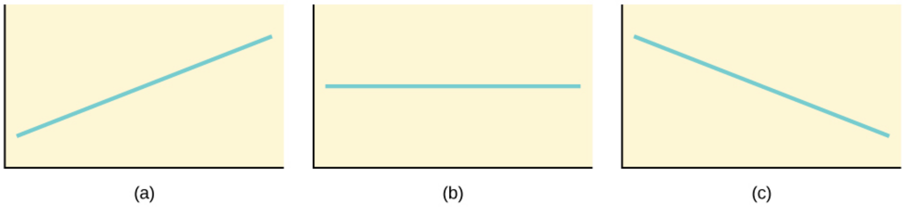
Emma’s Extreme Sports hires hang-gliding instructors and pays them a fee of $50 per class as well as $20 per student in the class. The total cost Emma pays depends on the number of students in a class. Find the equation that expresses the total cost in terms of the number of students in a class.

Solution:

$y = 50 + 20x$

Slope and *Y*-Intercept of a Linear Equation

For the linear equation $y = a + bx$, b = the slope and a = the y-coordinate of the y-intercept. From algebra recall that the **slope** is a number that describes the steepness of a line and the **y-intercept** is the point, where the line crosses the y-axis.



Three possible graphs of $y = a + bx$. (a) If $b > 0$, the line slopes upward to the right. (b) If $b = 0$, the line is horizontal. (c) If $b < 0$, the line slopes downward to the right.

The **slope of a line** is a value that describes the rate of change between the independent and dependent variables. The **slope** tells us how the dependent variable (y) changes for every one unit increase in the independent variable (x), on average. The **y-intercept** is used to describe the dependent variable when the independent variable equals zero.

Example:

Svetlana tutors to make extra money for college. For each tutoring session, she charges a one-time fee of $25 plus $15 per hour of tutoring. A linear equation that expresses the total amount of money Svetlana earns for each session she tutors is $y = 25 + 15x$.

Exercise:

Problem:

What are the independent and dependent variables? What is the y-intercept and what is the slope? Interpret them using complete sentences.

Solution:

The independent variable (x) is the number of hours Svetlana tutors each session. The dependent variable (y) is the amount, in dollars, Svetlana earns for each session.

The y-intercept is (0, 25) since a = 25. At the start of the tutoring session, Svetlana charges a one-time fee of $25 (this is when x = 0). The slope is 15 since b = 15. For each session, Svetlana earns $15 for each hour she tutors.

Note:
Try It
Exercise:

Problem:

Ethan repairs household appliances like dishwashers and refrigerators. For each visit, he charges $25 plus $20 per hour of work. A linear equation that expresses the total amount of money Ethan earns per visit is $y = 25 + 20x$.

What are the independent and dependent variables? What is the y-intercept and what is the slope? Interpret them using complete sentences.

Solution:

The independent variable (x) is the number of hours Ethan works each visit. The dependent variable (y) is the amount, in dollars, Ethan earns for each visit.

The y-intercept is (0, 25). At the start of a visit, Ethan charges a one-time fee of $25 (this is when x = 0). The slope is 20. For each visit, Ethan earns $20 for each hour he works.

References

Data from the Centers for Disease Control and Prevention.

Data from the National Center for HIV, STD, and TB Prevention.

Section Review

Algebraically, a linear equation typically takes the form $y = a + bx$, where a and b are constants, x is the independent variable, and y is the dependent variable. In the equation $y = a + bx$, the constant b that is multiplied by the x variable (b is called a coefficient) is known as the **slope**. The slope describes the rate of change between the independent and dependent variables; in other words, the rate of change describes the change that occurs in the dependent variable as the independent variable is changed. In the equation $y = a + bx$, the constant a is known as the y-coordinate of the y-intercept. Graphically, the y-intercept is the point where the graph of the line crosses the y-axis. At this point $x = 0$.

Note:
Caution
You may remember from algebra that a linear equation has the form $y = mx + b$ where m is the slope and b represents the y-coordinate of the y-intercept. With either form, the coefficient on the x variable represents the slope and the constant term gives the y-intercept.

Formula Review

$y = a + bx$ where a is the y-coordinate of the y-intercept and b is the slope. The variable x is the independent variable and y is the dependent variable.

Use the following information to answer the next three exercises. A vacation resort rents SCUBA equipment to certified divers. The resort charges an up-front fee of $25 and another fee of $12.50 an hour.
Exercise:

Problem: What are the dependent and independent variables?

Solution:

dependent variable: fee amount; independent variable: time

Exercise:

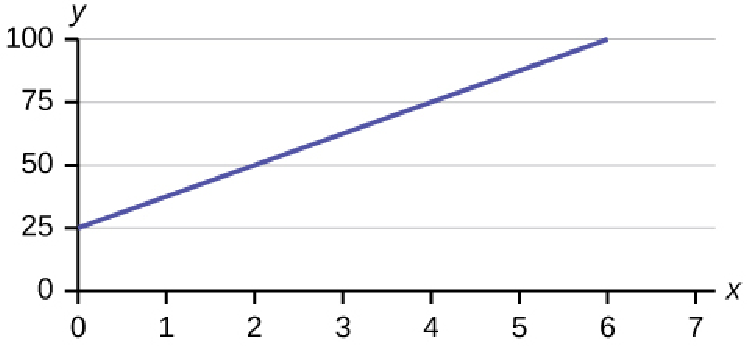
Problem:

Find the equation that expresses the total fee in terms of the number of hours the equipment is rented.

Exercise:

Problem: Graph the equation from [link].

Solution:



Use the following information to answer the next two exercises. A credit card company charges $10 when a payment is late, and $5 a day each day the payment remains unpaid.

Exercise:

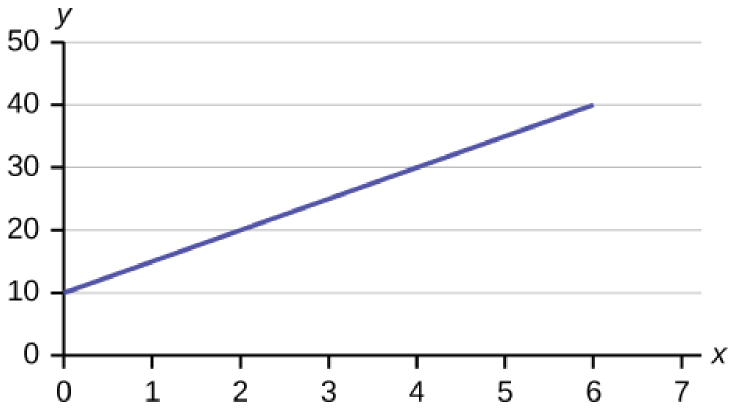
Problem:

Find the equation that expresses the total fee in terms of the number of days the payment is late.

Exercise:

Problem: Graph the equation from [link].

Solution:



Exercise:

Problem: Is the equation $y = 10 + 5x - 3x^2$ linear? Why or why not?

Exercise:

Problem: Which of the following equations are linear?

a. $y = 6x + 8$

b. $y + 7 = 3x$

c. $y - x = 8x^2$

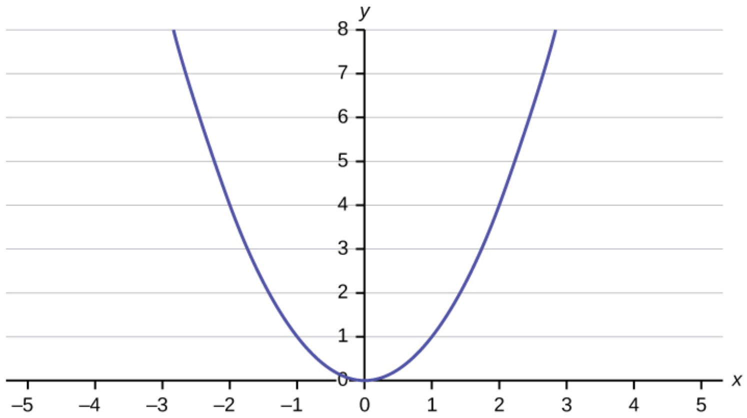
d. $4y = 8$

Solution:

a, b, and d are all linear equations.

Exercise:

Problem: Does the graph show a linear equation? Why or why not?



The following table contains real data for the first two decades of AIDS reporting.

Year	**# AIDS cases diagnosed**	**# AIDS deaths**
Pre-1981	91	29
1981	319	121
1982	1,170	453
1983	3,076	1,482
1984	6,240	3,466
1985	11,776	6,878
1986	19,032	11,987
1987	28,564	16,162

1988	35,447	20,868
1989	42,674	27,591
1990	48,634	31,335
1991	59,660	36,560
1992	78,530	41,055
1993	78,834	44,730
1994	71,874	49,095
1995	68,505	49,456
1996	59,347	38,510
1997	47,149	20,736
1998	38,393	19,005
1999	25,174	18,454
2000	25,522	17,347
2001	25,643	17,402
2002	26,464	16,371
Total	**802,118**	**489,093**

Adults and Adolescents only, United States

Exercise:

Problem:

Use the columns "year" and "# AIDS cases diagnosed. Why is “year” the independent variable and “# AIDS cases diagnosed.” the dependent variable (instead of the reverse)?

Solution:

The number of AIDS cases depends on the year. Therefore, year becomes the independent variable and the number of AIDS cases is the dependent variable.

Use the following information to answer the next two exercises. A specialty cleaning company charges an equipment fee and an hourly labor fee. A linear equation that expresses the total amount of the fee the company charges for each session is $y = 50 + 100x$.

Exercise:

Problem: What are the independent and dependent variables?

Exercise:

Problem:

What is the y-intercept and what is the slope? Interpret them using complete sentences.

Solution:

The y-intercept is (0, 50) ($a = 50$). At the start of the cleaning, the company charges a one-time fee of $50 (this is when $x = 0$). The slope is 100 ($b = 100$). For each session, the company charges $100 for each hour they clean.

Use the following information to answer the next three questions. Due to erosion, a river shoreline is losing several thousand pounds of soil each year. A linear equation that expresses the total amount of soil lost per year is $y = 12{,}000x$.

Exercise:

Problem: What are the independent and dependent variables?

Exercise:

Problem: How many pounds of soil does the shoreline lose in a year?

Solution:

12,000 pounds of soil

Exercise:

Problem: What is the y-intercept? Interpret its meaning.

Use the following information to answer the next two exercises. The price of a single issue of stock can fluctuate throughout the day. A linear equation that represents the price of stock for Shipment Express is $y = 15 – 1.5x$ where x is the number of hours passed in an eight-hour day of trading.

Exercise:

Problem: What are the slope and y-intercept? Interpret their meaning.

Solution:

The slope is –1.5 (b = –1.5). This means the stock is losing value at a rate of $1.50 per hour. The y-intercept is (0, 15) (a = 15). This means the price of stock before the trading day was $15.

Exercise:

Problem:

If you owned this stock, would you want a positive or negative slope? Why?

Homework

Exercise:

Problem:

For each of the following situations, state the independent variable and the dependent variable.

a. A study is done to determine if elderly drivers are involved in more motor vehicle fatalities than other drivers. The number of fatalities per 100,000 drivers is compared to the age of drivers.
b. A study is done to determine if the weekly grocery bill changes based on the number of family members.
c. Insurance companies base life insurance premiums partially on the age of the applicant.
d. Utility bills vary according to power consumption.
e. A study is done to determine if a higher education reduces the crime rate in a population.

Solution:

a. independent variable: age; dependent variable: fatalities
b. independent variable: # of family members; dependent variable: grocery bill
c. independent variable: age of applicant; dependent variable: insurance premium
d. independent variable: power consumption; dependent variable: utility
e. independent variable: higher education (years); dependent variable: crime rates

Exercise:

Problem:

Piece-rate systems are widely debated incentive payment plans. In a recent study of loan officer effectiveness, the following piece-rate system was examined:

% of goal reached	< 80	80	100	120
Incentive	n/a	\$4,000 with an additional \$125 added per percentage point from 81–99%	\$6,500 with an additional \$125 added per percentage point from 101–119%	\$9,500 with an additional \$125 added per percentage point starting at 121%

If a loan officer makes 95% of his or her goal, write the linear function that applies based on the incentive plan table. In context, explain the y-intercept and slope.

Glossary

Linear Equation

an equation of the form $y=a+bx$. Its graph forms a straight line.

Slope

a number that describes the steepness of a line, that is how much the dependent variable (y) changes for each unit the independent variable x increases.

y-intercept

the point, where the line crosses the y-axis. It's used to describe the dependent variable when the independent variable equals zero.

Scatterplots

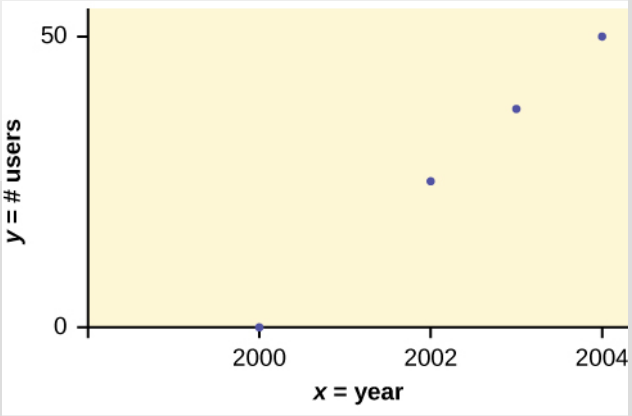
Before we take up the discussion of linear regression and correlation, we need to examine a way to display the relation between two variables x and y. The most common and easiest way is a **scatterplot**. The following example illustrates a scatterplot.

Example:
In Europe and Asia, m-commerce is popular. M-commerce users have special mobile phones that work like electronic wallets as well as provide phone and Internet services. Users can do everything from paying for parking to buying a TV set or soda from a machine to banking to checking sports scores on the Internet. For the years 2000 through 2004, was there a relationship between the year and the number of m-commerce users? Construct a scatter plot. Let x = the year and let y = the number of m-commerce users, in millions.

Table showing the number of m-commerce users (in millions) by year.

x **(year)**	y **(# of users)**
2000	0.5
2002	20.0
2003	33.0

Scatterplot showing the number of m-commerce users (in millions) by year.



x **(year)**	y **(# of users)**
2004	47.0

Note: To create a scatterplot on the calculator:

1. Enter your X data into list L1 and your Y data into list L2.
2. Press 2nd STATPLOT ENTER to use Plot 1. On the input screen for PLOT 1, highlight On and press ENTER. (Make sure the other plots are OFF.)
3. For TYPE: highlight the very first icon, which is the scatterplot, and press ENTER.
4. For Xlist:, enter L1 ENTER and for Ylist: L2 ENTER.
5. For Mark: it does not matter which symbol you highlight, but the square is the easiest to see. Press ENTER.
6. Make sure there are no other equations that could be plotted. Press Y = and clear any equations out.
7. Press the ZOOM key and then the number 9 (for menu item "ZoomStat") ; the calculator will fit the window to the data. You can press WINDOW to see the scaling of the axes.

Note:
Try It
Exercise:

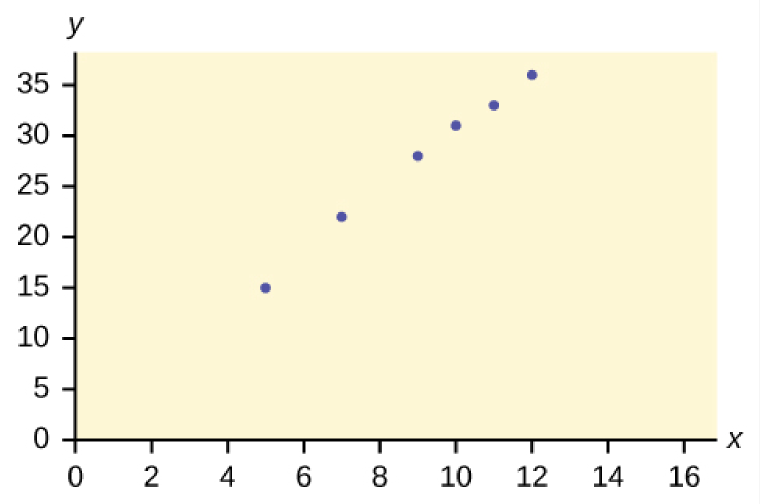
Problem:

Amelia plays basketball for her high school. She wants to improve to play at the college level. She notices that the number of points she scores in a game goes up in response to the number of hours she practices her jump shot each week. She records the following data:

X (hours practicing jump shot)	*Y* (points scored in a game)
5	15
7	22
9	28
10	31
11	33
12	36

Construct a scatterplot and state if what Amelia thinks appears to be true.

Solution:



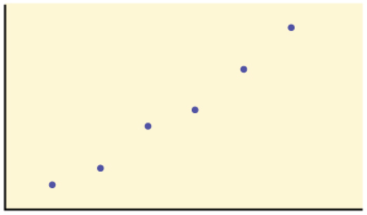
Yes, Amelia’s assumption appears to be correct. The number of points Amelia scores per game goes up when she practices her jump shot more.

A scatterplot shows the **direction** of a relationship between the variables. A clear direction happens when there is either:

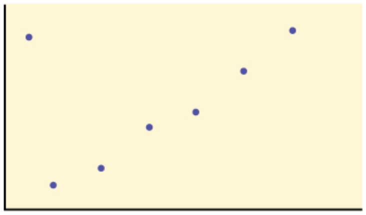
- High values of one variable occurring with high values of the other variable or low values of one variable occurring with low values of the other variable.
- High values of one variable occurring with low values of the other variable.

You can determine the **strength** of the relationship by looking at the scatterplot and seeing how close the points are to a line, a power function, an exponential function, or to some other type of function. For a linear relationship there is an exception. Consider a scatterplot where all the points fall on a horizontal line providing a "perfect fit." The horizontal line would in fact show no relationship.

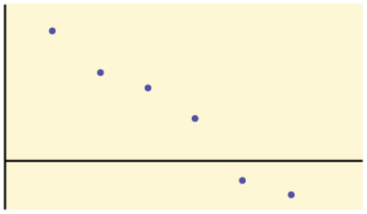
When you look at a scatterplot, you want to notice the **overall pattern** and any **deviations** from the pattern. The following scatterplot examples illustrate these concepts.



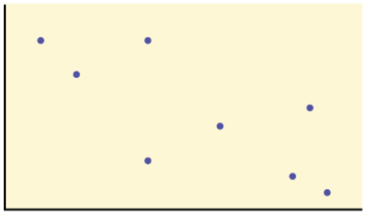
(a) Positive linear pattern (strong)



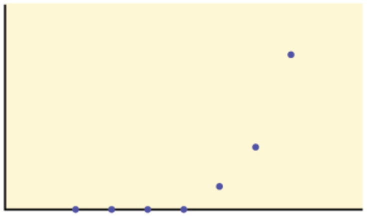
(b) Linear pattern w/ one deviation



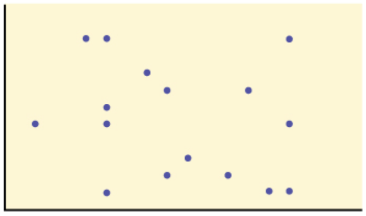
(a) Negative linear pattern (strong)



(b) Negative linear pattern (weak)



(a) Exponential growth pattern



(b) No pattern

In this chapter, we are interested in scatterplots that show a linear pattern. Linear patterns are quite common. The linear relationship is strong if the points are close to a straight line, except in the case of a horizontal line where there is no relationship. If we think that the points show a linear relationship, we would like to draw a line on the scatterplot. This line can be calculated through a process called **linear regression**. However, we only calculate a regression line if one of the variables helps to explain or predict the other variable. If x is the independent variable and y the dependent variable, then we can use a regression line to predict y for a given value of x.

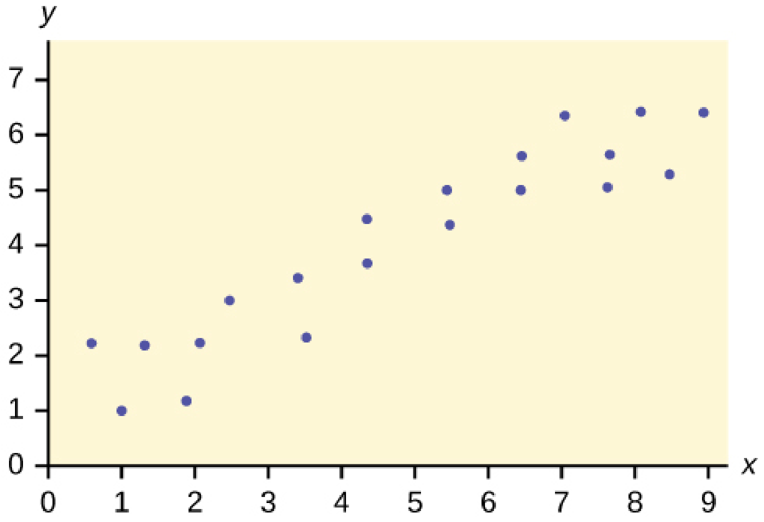
Section Review

Scatterplots are particularly helpful graphs when we want to see if there is a linear relationship among data points. They indicate both the direction of the relationship between the x variables and the y variables, and the strength of the relationship. We calculate the strength of the relationship between an independent variable and a dependent variable using linear regression.

Exercise:

Problem:

Does the scatterplot appear linear? Strong or weak? Positive or negative?



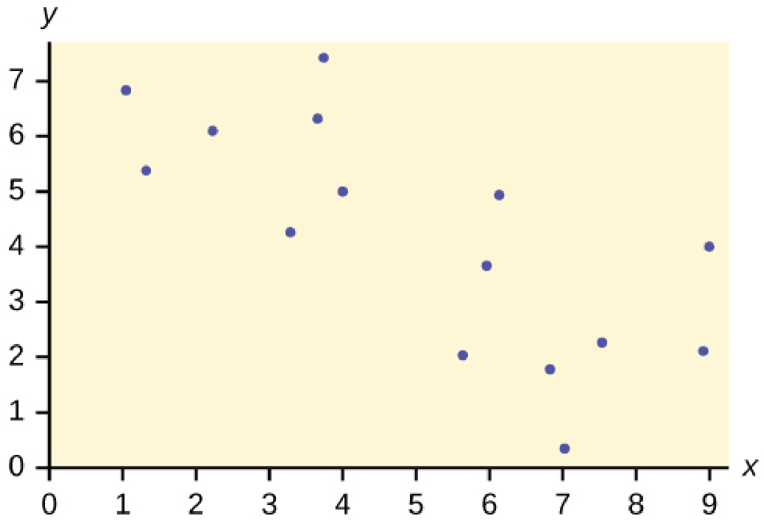
Solution:

The data appear to be linear with a strong, positive correlation.

Exercise:

Problem:

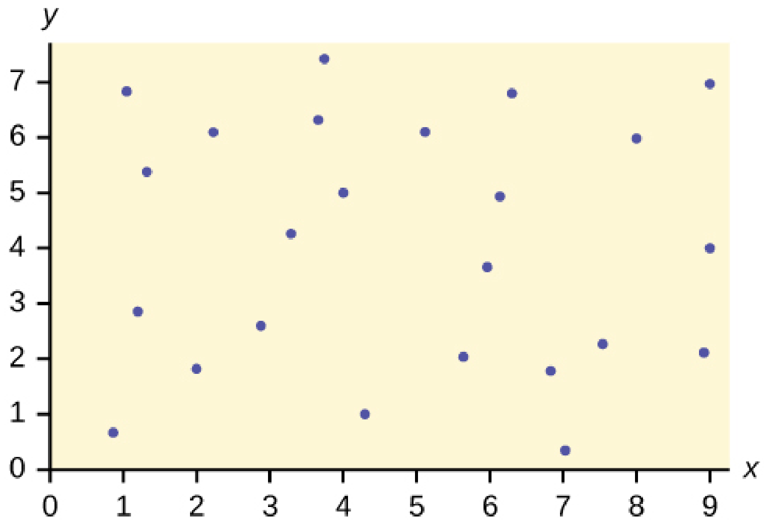
Does the scatterplot appear linear? Strong or weak? Positive or negative?



Exercise:

Problem:

Does the scatterplot appear linear? Strong or weak? Positive or negative?



Solution:

The data appear to have no correlation.

Homework

Exercise:

Problem:

The Gross Domestic Product Purchasing Power Parity is an indication of a country's currency value compared to another country. The following table shows the GDP PPP of Cuba as compared to US dollars. Construct a scatterplot of the data.

Year	Cuba's PPP	Year	Cuba's PPP
1999	1,700	2006	4,000

Year	Cuba’s PPP	Year	Cuba’s PPP
2000	1,700	2007	11,000
2002	2,300	2008	9,500
2003	2,900	2009	9,700
2004	3,000	2010	9,900
2005	3,500		

Solution:

Check student’s solution.

Exercise:

Problem:

The following table shows the poverty rates and cell phone usage in the United States. Construct a scatterplot of the data

Year	Poverty Rate	Cellular Usage per Capita
2003	12.7	54.67
2005	12.6	74.19
2007	12	84.86

Year	Poverty Rate	Cellular Usage per Capita
2009	12	90.82

Exercise:

Problem:

Does the higher cost of tuition translate into higher-paying jobs? The table lists the top ten colleges based on mid-career salary and the associated yearly tuition costs. Construct a scatterplot of the data.

School	Mid-Career Salary (in thousands)	Yearly Tuition
Princeton	137	28,540
Harvey Mudd	135	40,133
CalTech	127	39,900
US Naval Academy	122	0
West Point	120	0
MIT	118	42,050
Lehigh University	118	43,220

School	Mid-Career Salary (in thousands)	Yearly Tuition
NYU-Poly	117	39,565
Babson College	117	40,400
Stanford	114	54,506

Solution:

For graph: check student’s solution. Note that tuition is the independent variable and salary is the dependent variable.

The Regression Equation

Data rarely fit a straight line exactly. Usually, you must be satisfied with rough predictions. Typically, you have a set of data whose scatter plot appears to **"fit"** a straight line. This straight line is called the **Line of Best Fit or Least Squares Regression Line**.

Optional Collaborative Classroom Activity

If you know a person's pinky (smallest) finger length, do you think you could predict that person's height? Collect data from your class (pinky finger length, in inches). The independent variable, x, is pinky finger length and the dependent variable, y, is height.

For each set of data, plot the points on graph paper. Make your graph big enough and **use a ruler**. Then "by eye" draw a line that appears to "fit" the data. For your line, pick two convenient points and use them to find the slope of the line. Find the y-intercept of the line by extending your lines so they cross the y-axis. Using the slopes and the y-intercepts, write your equation of "best fit". Do you think everyone will have the same equation? Why or why not?

Using your equation, what is the predicted height for a pinky length of 2.5 inches?

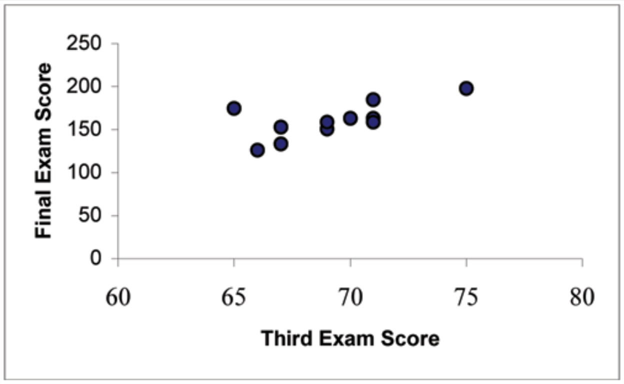
Example:
A random sample of 11 statistics students produced the following data where x is the third exam score, out of 80, and y is the final exam score, out of 200. Can you predict the final exam score of a random student if you know the third exam score?

Table showing the scores on the final exam based on scores

Scatterplot showing the scores on the final exam based on scores from the third exam.

x (third exam score)	y (final exam score)
65	175
67	133
71	185
71	163
66	126
75	198
67	153
70	163
71	159
69	151
69	159

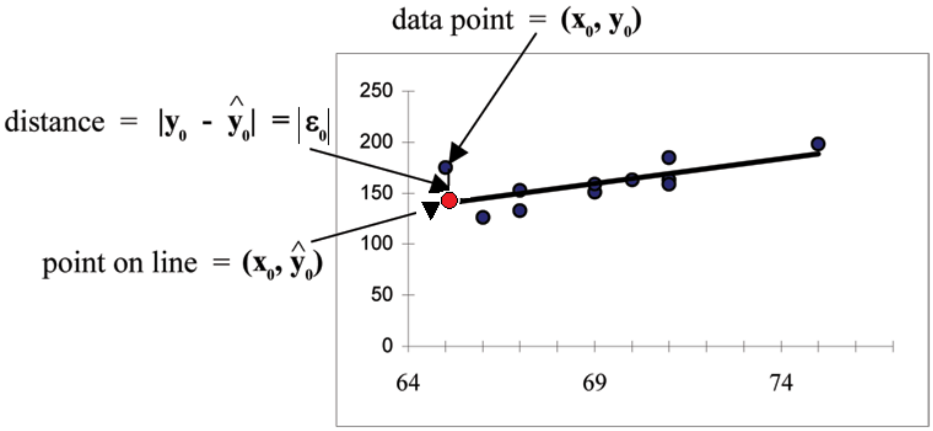
from the third exam.



The third exam score, x, is the independent variable and the final exam score, y, is the dependent variable. We will plot a regression line that best "fits" the data. If each of you were to fit a line "by eye", you would draw different lines. We can use what is called a **least-squares regression line** to obtain the best fit line.

Consider the following diagram. Each point of data is of the the form (x, y) and each point of the line of best fit using least-squares linear regression has the form $(x, \hat{y})$.

The $\hat{y}$ is read **"y hat"** and is the **estimated value of** y. It is the value of y obtained using the regression line. It is not generally equal to y from data.



The term $y_0 - \hat{y}_0 = \varepsilon_0$ is called the **"error" or residual**. It is not an error in the sense of a mistake. The **absolute value of a residual** measures the vertical distance between the actual value of y and the estimated value of y. In other words, it measures the vertical distance between the actual data point and the predicted point on the line.

ε = the Greek letter **epsilon**

If the observed data point lies above the line, the residual is positive, and the line underestimates the actual data value for y. If the observed data point lies below the line, the residual is negative, and the line overestimates that actual data value for y.

In the diagram above, $y_0 - \hat{y}_0 = \varepsilon_0$ is the residual for the point shown. Here the point lies above the line and the residual is positive.

For each data point, you can calculate the residuals or errors, $y_i - \hat{y}_i = \varepsilon_i$ for $i = 1$, 2, 3, ..., 11.

Each $|\varepsilon|$ is a vertical distance.

For the example about the third exam scores and the final exam scores for the 11 statistics students, there are 11 data points. Therefore, there are 11 ε values. If you square each ε and add, you get

$$\left(\varepsilon_1\right)^2 + \left(\varepsilon_2\right)^2 + ... + \left(\varepsilon_{11}\right)^2 = \sum_{\mathrm{i}=1}^{11} \varepsilon^2$$

This is called the **Sum of Squared Errors (SSE)**.

Using calculus, you can determine the values of a and b that make the **SSE** a minimum. When you make the **SSE** a minimum, you have determined the points that are on the line of best fit. It turns out that the line of best fit has the equation:

Equation:

$$\hat{y} = a + bx$$

where $a = \bar{y} - b \cdot \bar{x}$ and $b = \frac{\Sigma(x-\bar{x})\cdot(y-\bar{y})}{\Sigma(x-\bar{x})^2}$.

$\bar{x}$ and $\bar{y}$ are the sample means of the x values and the y values, respectively. The best fit line always passes through the point $(\bar{x}, \bar{y})$.

The slope b can be written as $b = r \cdot \left(\frac{s_y}{s_x}\right)$ where s_y = the standard deviation of the y values and s_x = the standard deviation of the x values. r is the correlation coefficient which is discussed in the next section.

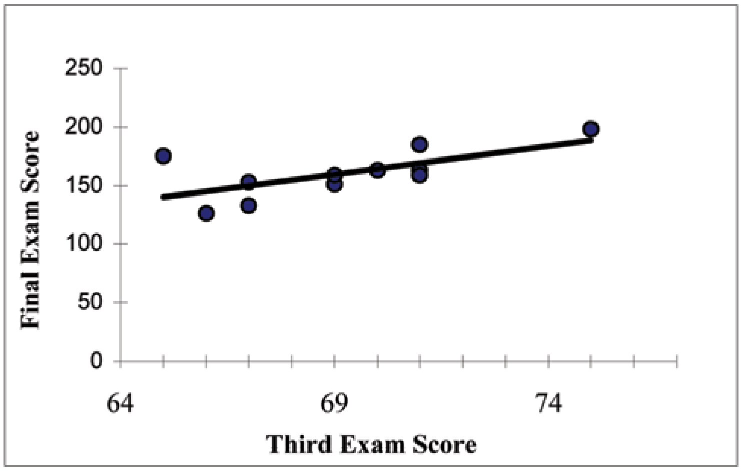
Least Squares Criteria for Best Fit

The process of fitting the best fit line is called **linear regression**. The idea behind finding the best fit line is based on the assumption that the data are scattered about a straight line. The criteria for the best fit line is that the sum of the squared errors (SSE) is minimized, that is made as small as possible. Any other line you might choose would have a higher SSE than the best fit line. This best fit line is called the **least squares regression line**
.

Note: Computer spreadsheets, statistical software, and many calculators can quickly calculate the best fit line and create the graphs. The calculations tend to be tedious if done by hand. Instructions to use the TI-83, TI-83+, and TI-84+ calculators to find the best fit line and create a scatterplot are shown at the end of this section.

THIRD EXAM vs FINAL EXAM EXAMPLE:

The graph of the line of best fit for the third exam/final exam example is shown below:



The least squares regression line (best fit line) for the third exam/final exam example has the equation:

Equation:

$$\hat{y} = -173.51 + 4.83x$$

Note:

- Remember, it is always important to plot a scatterplot first. If the scatterplot indicates that there is a linear relationship between the variables, then it is reasonable to use a best fit line to make predictions for y given x within the domain of x-values in the sample data, **but not necessarily for x-values outside that domain.**
- You could use the line to predict the final exam score for a student who earned a grade of 73 on the third exam.
- You should NOT use the line to predict the final exam score for a student who earned a grade of 50 on the third exam, because 50 is not within the domain of the x-values in the sample data, which are between 65 and 75.

UNDERSTANDING SLOPE
The slope of the line, b, describes how changes in the variables are related. It is important to interpret the slope of the line in the context of the situation represented by the data. You should be able to write a sentence interpreting the slope in plain English.

INTERPRETATION OF THE SLOPE: The slope of the best fit line tells us how the dependent variable (y) changes for every one unit increase in the independent (x) variable, on average.

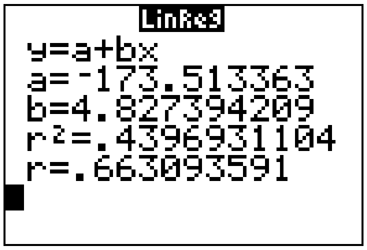
THIRD EXAM vs FINAL EXAM EXAMPLE

- Slope: The slope of the line is b = 4.83.
- Interpretation: For a one point increase in the score on the third exam, the final exam score increases by 4.83 points, on average.

Using the TI-83+ and TI-84+ Calculators

Using the calculator to find the linear regression equation

Turn on the diagnostics, by hitting 2nd Catalog (under the number 0), scroll to "Diagnostic On", hit enter and then enter again. (You only need to do this step once.)
In the STAT list editor, enter the X data in list L1 and the Y data in list L2, paired so that the corresponding (x,y) values are next to each other in the lists. (If a particular pair of values is repeated, enter it as many times as it appears in the data.)
On the STAT CALC menu, scroll down with the cursor to select the #8:LinReg(a+bx) and hit enter.
Hit enter one more time (or enter the appropriate list names, if requested, and then hit enter).



```
LinReg
y=a+bx
a=⁻173.513363
b=4.827394209
r²=.4396931104
r=.663093591
```

- The first line says y=a+bx. Scroll down to find the values a=-173.513363, and b=4.827394209 ; rounding each constant to two decimal places, the equation of the linear regression (best fit) line is $\hat{y} = -173.51 + 4.83x$
- The two items at the bottom are r^2 = .43969 and r=.663. For now, just note where to find these values; we will discuss them in the next two sections.

Graphing the Scatterplot and Regression Line

We are assuming your X data is already entered in list L1 and your Y data is in list L2
Press 2nd STATPLOT ENTER to use Plot 1
On the input screen for PLOT 1, highlight**On**and press ENTER
For TYPE: highlight the very first icon which is the scatterplot and press ENTER
Indicate Xlist: L1 and Ylist: L2
For Mark: it does not matter which symbol you highlight.
To graph the best fit line, press the "Y=" key and type the equation -173.51 + 4.83x into equation Y1. (The X key is immediately left of the STAT key).
Press the ZOOM key and then the number 9 (for menu item "ZoomStat"); the calculator will fit the window to the data and graph the line on the scatterplot
Optional: If you want to change the viewing window, press the WINDOW key. Enter your desired window using Xmin, Xmax, Ymin, Ymax

The Correlation Coefficient

Besides looking at the scatterplot and seeing that a line seems reasonable, how can you tell if the line is a good predictor? Use the correlation coefficient as another indicator (besides the scatterplot) of the strength of the relationship between x and y.

The **correlation coefficient, r,** developed by Karl Pearson in the early 1900s, is numerical and provides a measure of strength and direction of the linear association between the independent variable x and the dependent variable y.

The correlation coefficient is calculated as
Equation:

$$r = \frac{n\mathit{\Sigma}(xy) - (\mathit{\Sigma}x)(\mathit{\Sigma}y)}{\sqrt{\left[n\mathit{\Sigma}x^2 - (\mathit{\Sigma}x)^2\right]\left[n\mathit{\Sigma}y^2 - (\mathit{\Sigma}y)^2\right]}}$$

where n = the number of data points.

If you suspect, judging from the scatterplot, a linear relationship between x and y, then r can measure how strong the linear relationship is.

What the value of r tells us

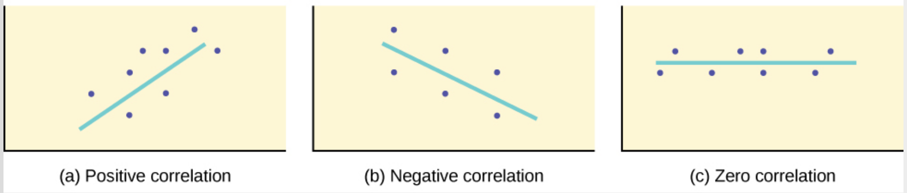
- The value of r is always between –1 and +1: $-1 \leq r \leq 1$.
- The size of the correlation r indicates the strength of the linear relationship between x and y. Values of r close to –1 or to +1 indicate a stronger linear relationship between x and y.
- If r = 0 there is absolutely no linear relationship between x and y **(no linear correlation)**.
- If r = 1, there is perfect positive correlation. If r = –1, there is perfect negative correlation. In both these cases, all of the original data points lie on a straight line. Of course,in the real world, this will not generally happen.

What the SIGN of r tells us

- A positive value of r means that when x increases, y tends to increase and when x decreases, y tends to decrease **(positive correlation)**.
- A negative value of r means that when x increases, y tends to decrease and when x decreases, y tends to increase **(negative correlation)**.
- The sign of r is the same as the sign of the slope, b, of the best-fit line.

Note: Strong correlation does not suggest that x causes y or y causes x. We say **"correlation does not imply causation."**

Note:



(a) A scatter plot showing data with a positive correlation. $0 < r < 1$ (b) A scatter plot showing data with a negative correlation. $-1 < r < 0$ (c) A scatter plot showing data with zero correlation. $r = 0$

The formula for r looks formidable. However, computer spreadsheets, statistical software, and many calculators can quickly calculate r. The correlation coefficient r is the second to last item in the output screens for the LinReg on the TI-83, TI-83+, or TI-84+ calculator (see previous section for instructions).

The Coefficient of Determination

The variable r^2 is called the coefficient of determination and is the square of the correlation coefficient, but is usually stated as a percent, rather than in decimal form. It has an interpretation in the context of the data:

- r^2, when expressed as a percent, represents the percent of variation in the dependent (predicted) variable y that can be explained by variation in the independent (explanatory) variable x using the regression (best-fit) line.
- $1 – r^2$, when expressed as a percentage, represents the percent of variation in y that is NOT explained by variation in x using the regression line. This can be seen as the scattering of the observed data points about the regression line.

Consider the third exam/final exam example introduced previously

- The line of best fit is: $\hat{y}$ = –173.51 + 4.83x
- The correlation coefficient is r = 0.6631
- The coefficient of determination is r^2 = 0.6631^2 = 0.4397
- **Interpretation of r^2 in the context of this example:**
- Approximately 44% of the variation (0.4397 is approximately 0.44) in the final-exam grades can be explained by the variation in the grades on the third exam, using the best-fit regression line.
- Therefore, approximately 56% of the variation (1 – 0.44 = 0.56) in the final exam grades can NOT be explained by the variation in the grades on the third exam, using the best-fit regression line. (This is seen as the scattering of the points about the line.)

Section Review

A regression line, or a line of best fit, can be drawn on a scatterplot and used to predict outcomes for the x and y variables in a given data set or sample data. There are several ways to find a regression line, but usually the least-squares regression line is used because it creates a uniform line. Residuals, also called “errors,” measure the distance from the actual value of y and the estimated value of y. The Sum of Squared Errors, when set to its minimum, calculates the points on the line of best fit. Regression lines can be used to predict values within the given set of data, but should not be used to make predictions for values outside the set of data.

The correlation coefficient r measures the strength of the linear association between x and y. The variable r has to be between –1 and +1. When r is positive, the x and y will tend to increase and decrease together. When r is negative, x will increase and y will decrease, or the opposite, x will decrease and y will increase. The coefficient of determination r^2, is equal to the square of the correlation coefficient. When expressed as a percent, r^2 represents the percent of variation in the dependent variable y that can be explained by variation in the independent variable x using the regression line.

Practice

Use the following information to answer the next five exercises. A random sample of ten professional athletes produced the following data where x is the number of endorsements the player has and y is the amount of money made (in millions of dollars).

x	***y***	***x***	***y***
0	2	5	12
3	8	4	9
2	7	3	9
1	3	0	3
5	13	4	10

Summary

Exercise:

Problem: Draw a scatterplot of the data.

Exercise:

Problem: Use regression to find the equation for the line of best fit.

Solution:

$\hat{y}$ = 2.23 + 1.99x

Exercise:

Problem: Draw the line of best fit on the scatterplot.

Exercise:

Problem:

What is the slope of the line of best fit? What does it represent?

Solution:

The slope is 1.99 (b = 1.99). It means that for every endorsement deal a professional player gets, he gets an average of another $1.99 million in pay each year.

Exercise:

Problem:

What is the y-intercept of the line of best fit? What does it represent?

Exercise:

Problem: What does an r value of zero mean?

Solution:

It means that there is no correlation between the data sets.

Exercise:

Problem:

When n = 100 and r = -0.89, is there a significant correlation? Explain.

Solution:

Yes, there are enough data points and the value of r is strong enough to show that there is a strong negative correlation between the data sets.

Homework

Exercise:

Problem:

What is the process through which we can calculate a line that goes through a scatterplot with a linear pattern?

Exercise:

Problem: Explain what it means when a correlation has an r^2 of 0.72.

Solution:

It means that 72% of the variation in the dependent variable (y) can be explained by the variation in the independent variable (x).

Exercise:

Problem:

Can a coefficient of determination be negative? Why or why not?

Glossary

Line of Best Fit or Least Squares Regression Line

the straight line which best fits a set of data points on a scatterplot by minimizing the sum of the squared errors (SSE).

Linear Regression

the process of finding the line of best fit or least squares regression line.

Coefficient of Correlation

a measure developed by Karl Pearson (early 1900s) that gives the strength of association between the independent variable and the dependent variable; the formula is:

Equation:

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

where n is the number of data points. The coefficient cannot be more then 1 and less then –1. The closer the coefficient is to ±1, the stronger the evidence of a significant linear relationship between x and y.

Coefficient of Determination

the square of the correlation coefficient, usually stated as a percent, and represents the percent of variation in the dependent (predicted) variable y that can be explained by variation in the independent (explanatory) variable x using the regression (best-fit) line.

Prediction

Recall the third exam/final exam example.

We examined the scatterplot and showed that the correlation coefficient is significant. We found the equation of the best-fit line for the final exam grade as a function of the grade on the third-exam. We can now use the least-squares regression line for prediction.

Suppose you want to estimate, or predict, the mean final exam score of statistics students who received 73 on the third exam. The exam scores **(x-values)** range from 65 to 75. **Since 73 is between the x-values 65 and 75**, substitute x = 73 into the equation. Then:

Equation:

$$\hat{y} = -173.51 + 4.83(73) = 179.08$$

We predict that statistics students who earn a grade of 73 on the third exam will earn a grade of 179.08 on the final exam, on average.

Example:
Recall the third exam/final exam example.

Exercise:

Problem:

a. What would you predict the final exam score to be for a student who scored a 66 on the third exam?

Solution:

a. 145.27

Exercise:

Problem:

b. What would you predict the final exam score to be for a student who scored a 90 on the third exam?

Solution:

b. The x values in the data are between 65 and 75. Ninety is outside of the domain of the observed x values in the data (independent variable), so you cannot reliably predict the final exam score for this student. (Even though it is possible to enter 90 into the equation for x and calculate a corresponding y value, the y value that you get will not be reliable.)

To understand really how unreliable the prediction can be outside of the observed x values observed in the data, make the substitution x = 90 into the equation.

$$\hat{y} = -173.51 + 4.83(90) = 261.19$$

The final-exam score is predicted to be 261.19. The largest the final-exam score can be is 200.

Note: Note
The process of predicting inside of the observed x values observed in the data is called **interpolation**. The process of predicting outside of the observed x values observed in the data is called **extrapolation**.

Note:
Try It
Exercise:

Problem:

Data are collected on the relationship between the number of hours per week practicing a musical instrument and scores on a math test. The line of best fit is as follows:

$\hat{y}$ = 72.5 + 2.8x

What would you predict the score on a math test would be for a student who practices a musical instrument for five hours a week?

Solution:

86.5

References

Data from the Centers for Disease Control and Prevention.

Data from the National Center for HIV, STD, and TB Prevention.

Data from the United States Census Bureau. Available online at http://www.census.gov/compendia/statab/cats/transportation/motor_vehicle_accidents_and_fatalities.html

Data from the National Center for Health Statistics.

Section Review

After determining the presence of a strong correlation coefficient and calculating the line of best fit, you can use the least squares regression line to make predictions about your data.

Use the following information to answer the next two exercises. An electronics retailer used regression to find a simple model to predict sales growth in the first quarter of the new year (January through March). The model is good for 90 days, where x is the day. The model can be written as follows:

$\hat{y}$ = 101.32 + 2.48x where $\hat{y}$ is in thousands of dollars.

Exercise:

Problem: What would you predict the sales to be on day 60?

Solution:

$250,120

Exercise:

Problem: What would you predict the sales to be on day 90?

Use the following information to answer the next three exercises. A landscaping company is hired to mow the grass for several large properties. The total area of the properties combined is 1,345 acres. The rate at which one person can mow is as follows:

$\hat{y}$ = 1350 – 1.2x where x is the number of hours and $\hat{y}$ represents the number of acres left to mow.

Exercise:

Problem: How many acres will be left to mow after 20 hours of work?

Solution:

1,326 acres

Exercise:

Problem:

How many acres will be left to mow after 100 hours of work?

Exercise:

Problem:

How many hours will it take to mow all of the lawns? (When is $\hat{y} = 0$?)

Solution:

1,125 hours, or when $x = 1{,}125$

The following table contains real data for the first two decades of AIDS reporting.

Year	# AIDS cases diagnosed	# AIDS deaths
Pre-1981	91	29
1981	319	121
1982	1,170	453
1983	3,076	1,482
1984	6,240	3,466
1985	11,776	6,878
1986	19,032	11,987

1987	28,564	16,162
1988	35,447	20,868
1989	42,674	27,591
1990	48,634	31,335
1991	59,660	36,560
1992	78,530	41,055
1993	78,834	44,730
1994	71,874	49,095
1995	68,505	49,456
1996	59,347	38,510
1997	47,149	20,736
1998	38,393	19,005
1999	25,174	18,454
2000	25,522	17,347
2001	25,643	17,402
2002	26,464	16,371
Total	**802,118**	**489,093**

Adults and Adolescents only, United States

Exercise:

Problem:

Graph “year” versus “# AIDS cases diagnosed” (plot the scatter plot). Do not include pre-1981 data.

Exercise:

Problem:

Perform linear regression. What is the linear equation? Round each constant to the nearest whole number.

Solution:

Check student’s solution.

Exercise:

Problem: Write the equations:

a. Linear equation: ___________
b. a = _________
c. b = _________
d. r = _________
e. n = _________

Exercise:

Problem: Solve.

a. When x = 1985, $\hat{y}$ = ______
b. When x = 1990, $\hat{y}$ =______
c. When x = 1970, $\hat{y}$ =_______ Why doesn’t this answer make sense?

Solution:

a. When x = 1985, $\hat{y}$ = 25,525
b. When x = 1990, $\hat{y}$ = 34,275
c. When x = 1970, $\hat{y}$ = –725. Why doesn't this answer make sense? The range of x values was 1981 to 2002; the year 1970 is not in this range. The regression equation does not apply, because predicting for the year 1970 is extrapolation, which requires a different process. Also, a negative number does not make sense in this context, where we are predicting AIDS cases diagnosed.

Exercise:

Problem: Does the line seem to fit the data? Why or why not?

Exercise:

Problem:

Plot the two given points (1985, 25525) and (1990, 34275) on the following graph. Then, connect the two points to form the regression line.

Obtain the graph on your calculator or computer.

Exercise:

Problem: Write the equation: $\hat{y}$ = _____________

Solution:

$\hat{y} = 1750x – 3448225$

Exercise:

Problem:

Hand draw a smooth curve on the graph that shows the flow of the data.

Exercise:

Problem: Does the line seem to fit the data? Why or why not?

Solution:

There was an increase in AIDS cases diagnosed until 1993. From 1993 through 2002, the number of AIDS cases diagnosed declined each year. It is not appropriate to use a linear regression line to fit to the data.

Exercise:

Problem: Do you think a linear fit is best? Why or why not?

Exercise:

Problem:

What does the correlation imply about the relationship between time (years) and the number of diagnosed AIDS cases reported in the U.S.?

Solution:

Since there is no linear association between year and # of AIDS cases diagnosed, it is not appropriate to calculate a linear correlation coefficient. Even when there is a linear association and it is appropriate to calculate a correlation, we cannot say that one variable “causes” the other variable.

Exercise:

Problem:

Graph “year” vs. “# AIDS cases diagnosed.” Do not include pre-1981. Label both axes with words. Scale both axes.

Exercise:

Problem:

Enter your data into your calculator or computer. The pre-1981 data should not be included. Why is that so?

Write the linear equation, rounding the slope to four decimal places:

Solution:

We don’t know if the pre-1981 data was collected from a single year. So we don’t have an accurate x value for this figure.

Regression equation: $\hat{y}$ (#AIDS Cases) = 1749.777x (year) – 3,448,225

Exercise:

Problem: Calculate the following:

a. a = ______
b. b = ______
c. correlation coefficient = ______
d. n = ______

Homework

Exercise:

Problem:

Recently, the annual number of driver deaths per 100,000 for the selected age groups was as follows:

Age	**Number of Driver Deaths per 100,000**
16–19	38
20–24	36
25–34	24
35–54	20
55–74	18
75+	28

a. For each age group, pick the midpoint of the interval for the x value. (For the 75+ group, use 80.)
b. Using "ages" as the independent variable and "Number of driver deaths per 100,000" as the dependent variable, make a scatterplot of the data.
c. Calculate the least squares (best–fit) line. Put the equation in the form of: $\hat{y} = a + bx$
d. Find the correlation coefficient.
e. Predict the number of deaths for ages 40 and 60.
f. Based on the given data, does there appear to be a linear relationship between age of a driver and driver fatality rate?
g. What is the slope of the least squares (best-fit) line? Interpret the slope.

Solution:

a.

Age	**Number of Driver Deaths per 100,000**
17.5	38
22	36
29.5	24
44.5	20
64.5	18
80	28

b. Check student’s solution.
c. $\hat{y}$ = 35.5818045 – 0.19182491x
d. r = –0.57874
e. If age = 40, $\hat{y}$ (deaths) = 35.5818045 – 0.19182491(40) = 27.9.

If age = 60, $\hat{y}$ (deaths) = 35.5818045 – 0.19182491(60) = 24.1.
f. Based on this data, there appears to be a linear relationship for the ages up to age 74. The oldest age group shows an increase in deaths from the prior group, which is not consistent with the younger ages.
g. slope = –0.19182491. For each year of age, the number of driver deaths decreases by 0.19182491 per 100,000.

Exercise:

Problem:

[link] shows the life expectancy for an individual born in the United States in certain years.

Year of Birth	Life Expectancy
1930	59.7
1940	62.9
1950	70.2
1965	69.7
1973	71.4
1982	74.5
1987	75
1992	75.7
2010	78.7

a. Decide which variable should be the independent variable and which should be the dependent variable.
b. Draw a scatterplot of the ordered pairs.
c. Calculate the least squares line. Put the equation in the form of: $\hat{y} = a + bx$
d. Find the correlation coefficient.

e. Find the estimated life expectancy for an individual born in 1950 and for one born in 1982.
f. Why aren’t the answers to part e the same as the values in [link] that correspond to those years?
g. Use the two points in part e to plot the least squares line on your graph from part b.
h. Based on the data, does there appear to be a linear relationship between the year of birth and life expectancy?
i. Using the least squares line, find the estimated life expectancy for an individual born in 1850. Does the least squares line give an accurate estimate for that year? Explain why or why not.
j. What is the slope of the least-squares (best-fit) line? Interpret the slope.

Exercise:

Problem:

The maximum discount value of the Entertainment® card for the “Fine Dining” section, Edition ten, for various pages is given in [link]

Page number	Maximum value ($)
4	16
14	19
25	15
32	17
43	19

Page number	**Maximum value ($)**
57	15
72	16
85	15
90	17

a. Decide which variable should be the independent variable and which should be the dependent variable.
b. Draw a scatterplot of the ordered pairs.
c. Calculate the least-squares line. Put the equation in the form of: $\hat{y} = a + bx$
d. Find the correlation coefficient.
e. Find the estimated maximum values for the restaurants on page ten and on page 70.
f. Does it appear that the restaurants giving the maximum value are placed in the beginning of the "Fine Dining" section? How did you arrive at your answer?
g. Suppose that there were 200 pages of restaurants. What do you estimate to be the maximum value for a restaurant listed on page 200?
h. Is the least squares line valid for page 200? Why or why not?
i. What is the slope of the least-squares (best-fit) line? Interpret the slope.

Solution:

a. We wonder if the better discounts appear earlier in the book so we select page as X and discount as Y.
b. Check student's solution.
c. $\hat{y} = 17.21757 – 0.01412x$
d. $r = – 0.2752$

e. page 10: 17.08 page 70: 16.23
f. Based on this data, it appears there is no relationship between the page and the amount of the discount.
g. page 200: 14.39
h. No, using the regression equation to predict for page 200 is extrapolation.
i. slope = –0.01412. As the page number increases by one page, the discount decreases by $0.01412.

Exercise:

Problem:

[link] gives the gold medal times for every other Summer Olympics for the women's 100-meter freestyle (swimming).

Year	Time (seconds)
1912	82.2
1924	72.4
1932	66.8
1952	66.8
1960	61.2
1968	60.0
1976	55.65

Year	Time (seconds)
1984	55.92
1992	54.64
2000	53.8
2008	53.1

a. Decide which variable should be the independent variable and which should be the dependent variable.
b. Draw a scatterplot of the data.
c. Does it appear from inspection that there is a relationship between the variables? Why or why not?
d. Calculate the least squares line. Put the equation in the form of: $\hat{y} = a + bx$.
e. Find the correlation coefficient.
f. Find the estimated gold medal time for 1932. Find the estimated time for 1984.
g. Why are the answers from part f different from the chart values?
h. Does it appear that a line is the best way to fit the data? Why or why not?
i. Use the least-squares line to estimate the gold medal time for the next Summer Olympics. Do you think that your answer is reasonable? Why or why not?

Exercise:

Problem:

State	# letters in name	Year entered the Union	Rank for entering the Union	Area (square miles)
Alabama	7	1819	22	52,423
Colorado	8	1876	38	104,100
Hawaii	6	1959	50	10,932
Iowa	4	1846	29	56,276
Maryland	8	1788	7	12,407
Missouri	8	1821	24	69,709
New Jersey	9	1787	3	8,722
Ohio	4	1803	17	44,828
South Carolina	13	1788	8	32,008
Utah	4	1896	45	84,904
Wisconsin	9	1848	30	65,499

We are interested in whether or not the number of letters in a state name depends upon the year the state entered the Union.

a. Decide which variable should be the independent variable and which should be the dependent variable.
b. Draw a scatterplot of the data.

c. Does it appear from inspection that there is a relationship between the variables? Why or why not?
d. Calculate the least-squares line. Put the equation in the form of: $\hat{y} = a + bx$.
e. Find the correlation coefficient. What does it imply about the strength of the linear relationship?
f. Find the estimated number of letters (to the nearest integer) a state would have if it entered the Union in 1900. Find the estimated number of letters a state would have if it entered the Union in 1940.
g. Does it appear that a line is the best way to fit the data? Why or why not?
h. Use the least-squares line to estimate the number of letters a new state that enters the Union the year 2013 would have. Can the least squares line be used to predict it? Why or why not?

Solution:

a. Year is the independent or x variable; the number of letters is the dependent or y variable.
b. Check student's solution.
c. no
d. $\hat{y} = 47.03 – 0.0216x$
e. –0.4280
f. 6; 5
g. No, the relationship does not appear to be linear.
h. 3.55 or four letters; this is not an appropriate use of the least squares line. It is extrapolation.

Testing the Significance of the Correlation Coefficient (Optional)

The correlation coefficient, r, tells us about the strength and direction of the linear relationship between x and y. However, the reliability of the linear model also depends on how many observed data points are in the sample. We need to look at both the value of the correlation coefficient r and the sample size n, together.

We perform a hypothesis test of the **"significance of the correlation coefficient"** to decide whether the linear relationship in the sample data is strong enough to use to model the relationship in the population.

The sample data are used to compute r, the correlation coefficient for the sample. If we had data for the entire population, we could find the population correlation coefficient. But because we have only have sample data, we cannot calculate the population correlation coefficient. The sample correlation coefficient, r, is our estimate of the unknown population correlation coefficient.

- The symbol for the population correlation coefficient is ρ, the Greek letter "rho."
- ρ = population correlation coefficient (unknown)
- r = sample correlation coefficient (known; calculated from sample data)

The hypothesis test lets us decide whether the value of the population correlation coefficient ρ is "close to zero" or "significantly different from zero". We decide this based on the sample correlation coefficient r and the sample size n.

If the test concludes that the correlation coefficient is significantly different from zero, we say that the correlation coefficient is "significant."

- Conclusion: There is sufficient evidence to conclude that there is a significant linear relationship between x and y because the correlation coefficient is significantly different from zero.

- What the conclusion means: There is a significant linear relationship between x and y. We can use the regression line to model the linear relationship between x and y in the population.

If the test concludes that the correlation coefficient is not significantly different from zero (it is close to zero), we say that correlation coefficient is "not significant".

- Conclusion: "There is insufficient evidence to conclude that there is a significant linear relationship between x and y because the correlation coefficient is not significantly different from zero."
- What the conclusion means: There is not a significant linear relationship between x and y. Therefore, we CANNOT use the regression line to model a linear relationship between x and y in the population.

Note:
Note

- If r is significant and the scatter plot shows a linear trend, the line can be used to predict the value of y for values of x that are within the domain of observed x values.
- If r is not significant OR if the scatter plot does not show a linear trend, the line should not be used for prediction.
- If r is significant and if the scatter plot shows a linear trend, the line may NOT be appropriate or reliable for prediction OUTSIDE the domain of observed x values in the data.

PERFORMING THE HYPOTHESIS TEST

- **Null Hypothesis: H_0: $\rho = 0$**
- **Alternate Hypothesis: H_a: $\rho \neq 0$**

WHAT THE HYPOTHESES MEAN IN WORDS:

- **Null Hypothesis H_0:** The population correlation coefficient IS NOT significantly different from zero. There IS NOT a significant linear relationship (correlation) between x and y in the population.
- **Alternate Hypothesis H_a:** The population correlation coefficient IS significantly DIFFERENT FROM zero. There IS A SIGNIFICANT LINEAR RELATIONSHIP (correlation) between x and y in the population.

DRAWING A CONCLUSION:
There are two methods of making the decision. The two methods are equivalent and give the same result.

- **Method 1: Using the p-value**
- **Method 2: Using a table of critical values**

In this chapter of this textbook, we will always use a significance level of 5%, $\alpha = 0.05$

Note:
Note
Using the p-value method, you could choose any appropriate significance level you want; you are not limited to using $\alpha = 0.05$. But the table of critical values provided in this textbook assumes that we are using a significance level of 5%, $\alpha = 0.05$. (If we wanted to use a different significance level than 5% with the critical value method, we would need different tables of critical values that are not provided in this textbook.)

METHOD 1: Using a p-value to make a decision

Note:

To calculate the p-value using LinRegTTEST:
On the LinRegTTEST input screen, on the line prompt for β or ρ, highlight "**≠ 0**"
The output screen shows the p-value on the line that reads "p =".
(Most computer statistical software can calculate the p-value.)

If the p-value is less than the significance level (α = 0.05):

- Decision: Reject the null hypothesis.
- Conclusion: "There is sufficient evidence to conclude that there is a significant linear relationship between x and y because the correlation coefficient is significantly different from zero."

If the p-value is NOT less than the significance level (α = 0.05)

- Decision: DO NOT REJECT the null hypothesis.
- Conclusion: "There is insufficient evidence to conclude that there is a significant linear relationship between x and y because the correlation coefficient is NOT significantly different from zero."

Calculation Notes:

- You will use technology to calculate the p-value. The following describes the calculations to compute the test statistics and the p-value:
- The p-value is calculated using a t-distribution with $n-$ 2 degrees of freedom.
- The formula for the test statistic is $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$. The value of the test statistic, t, is shown in the computer or calculator output along with the p-value. The test statistic t has the same sign as the correlation coefficient r.
- The p-value is the combined area in both tails.

An alternative way to calculate the p-value p given by LinRegTTest is the command 2*tcdf(abs(t),10^99, n-2) in 2nd DISTR.

THIRD-EXAM vs FINAL-EXAM EXAMPLE: p>-value method

- Consider the third exam/final exam example.
- The line of best fit is: $\hat{y}$ = -173.51 + 4.83x with r = 0.6631 and there are n = 11 data points.
- Can the regression line be used for prediction? **Given a third exam score (x value), can we use the line to predict the final exam score (predicted y value)?**

H_0: $\rho = 0$

H_a: $\rho \neq 0$

$\alpha = 0.05$

- The p-value is 0.026 (from LinRegTTest on your calculator or from computer software).
- The p-value, 0.026, is less than the significance level of α = 0.05.
- Decision: Reject the Null Hypothesis H_0
- Conclusion: There is sufficient evidence to conclude that there is a significant linear relationship between the third exam score (x) and the final exam score (y) because the correlation coefficient is significantly different from zero.

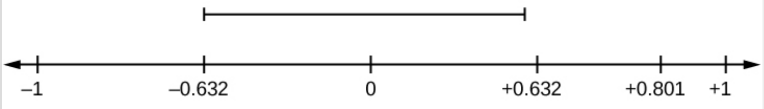
Because r is significant and the scatterplot shows a linear trend, the regression line can be used to predict final exam scores.

METHOD 2: Using a table of Critical Values to make a decision

The 95% Critical Values of the Sample Correlation Coefficient Table can be used to give you a good idea of whether the computed value of r is significant or not. Compare r to the appropriate critical value in the table. If r is not between the positive and negative critical values, then the correlation coefficient is significant. If r is significant, then you may want to use the line for prediction.

Example:

Suppose you computed r = 0.801 using n = 10 data points. $df = n - 2$ = 10 - 2 = 8. The critical values associated with df = 8 are -0.632 and +0.632. If r < negative critical value or r > positive critical value, then r is significant. Since r = 0.801 and 0.801 > 0.632, r is significant and the line may be used for prediction. If you view this example on a number line, it will help you.



r is not significant between -0.632 and +0.632. r = 0.801 > +0.632. Therefore, r is significant.

Note:

Try It

Exercise:

Problem:

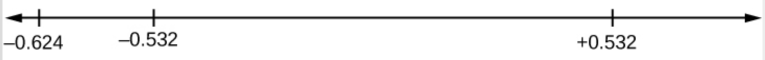
For a given line of best fit, you computed that r = 0.6501 using n = 12 data points and the critical value is 0.576. Can the line be used for prediction? Why or why not?

Solution:

If the scatterplot looks linear then, yes, the line can be used for prediction, because r > the positive critical value.

Example:

Suppose you computed r = –0.624 with 14 data points. df = 14 – 2 = 12. The critical values are –0.532 and 0.532. Since –0.624 < –0.532, r is significant and the line can be used for prediction



r = –0.624 < –0.532. Therefore, r is significant.

Note:
Try It
Exercise:

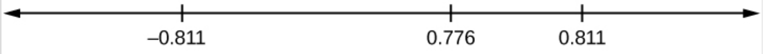
Problem:

For a given line of best fit, you compute that r = 0.5204 using n = 9 data points, and the critical value is 0.666. Can the line be used for prediction? Why or why not?

Solution:

No, the line cannot be used for prediction, because r < the positive critical value.

Example:
Suppose you computed r = 0.776 and n = 6. df = 6 – 2 = 4. The critical values are –0.811 and 0.811. Since –0.811 < 0.776 < 0.811, r is not significant, and the line should not be used for prediction.



-0.811 < r = 0.776 < 0.811. Therefore, r is not significant.

Note:
Try It
Exercise:

Problem:

For a given line of best fit, you compute that r = –0.7204 using n = 8 data points, and the critical value is 0.707. Can the line be used for prediction? Why or why not?

Solution:

Yes, the line can be used for prediction, because r < the negative critical value.

THIRD-EXAM vs FINAL-EXAM EXAMPLE: critical value method

Consider the third exam/final exam example. The line of best fit is: $\hat{y}$ = –173.51+4.83x with r = 0.6631 and there are n = 11 data points. Can the regression line be used for prediction? **Given a third-exam score (x value), can we use the line to predict the final exam score (predicted y value)?**

- H_0: $\rho = 0$
- H_a: $\rho \neq 0$
- $\alpha = 0.05$

- Use the "95% Critical Value" table for r with df = n – 2 = 11 – 2 = 9.
- The critical values are –0.602 and +0.602
- Since 0.6631 > 0.602, r is significant.

- Decision: Reject the null hypothesis.
- Conclusion:There is sufficient evidence to conclude that there is a significant linear relationship between the third exam score (x) and the final exam score (y) because the correlation coefficient is significantly different from zero.

Because r is significant and the scatter plot shows a linear trend, the regression line can be used to predict final exam scores.

Example:
Suppose you computed the following correlation coefficients. Using the table at the end of the chapter, determine if r is significant and the line of best fit associated with each r can be used to predict a y value. If it helps, draw a number line.

a. r = –0.567 and the sample size, n, is 19. The df = n – 2 = 17. The critical value is –0.456. –0.567 < –0.456 so r is significant.
b. r = 0.708 and the sample size, df, is nine. The n = n – 2 = 7. The critical value is 0.666. 0.708 > 0.666 so r is significant.
c. r = 0.134 and the sample size, n, is 14. The df = 14 – 2 = 12. The critical value is 0.532. 0.134 is between –0.532 and 0.532 so r is not significant.
d. r = 0 and the sample size, n, is five. No matter what the dfs are, r = 0 is between the two critical values so r is not significant.

Note:
Try It
Exercise:

Problem:

For a given line of best fit, you compute that r = 0 using n = 100 data points. Can the line be used for prediction? Why or why not?

Solution:

No, the line cannot be used for prediction no matter what the sample size is.

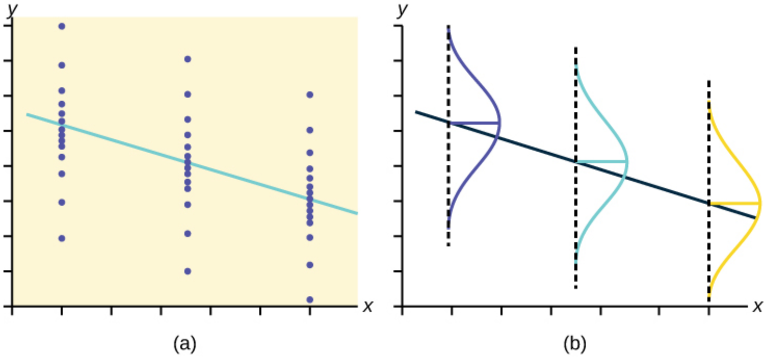
Assumptions in Testing the Significance of the Correlation Coefficient

Testing the significance of the correlation coefficient requires that certain assumptions about the data are satisfied. The premise of this test is that the data are a sample of observed points taken from a larger population. We have not examined the entire population because it is not possible or feasible to do so. We are examining the sample to draw a conclusion about whether the linear relationship that we see between x and y in the sample data provides strong enough evidence so that we can conclude that there is a linear relationship between x and y in the population.

The regression line equation that we calculate from the sample data gives the best-fit line for our particular sample. We want to use this best-fit line for the sample as an estimate of the best-fit line for the population. Examining the scatterplot and testing the significance of the correlation coefficient helps us determine if it is appropriate to do this.
The assumptions underlying the test of significance are:

- There is a linear relationship in the population that models the average value of y for varying values of x. In other words, the expected value of y for each particular value lies on a straight line in the population. (We do not know the equation for the line for the population. Our regression line from the sample is our best estimate of this line in the population.)
- The y values for any particular x value are normally distributed about the line. This implies that there are more y values scattered closer to the line than are scattered farther away. Assumption (1) implies that these normal distributions are centered on the line: the means of these normal distributions of y values lie on the line.

- The standard deviations of the population y values about the line are equal for each value of x. In other words, each of these normal distributions of y values has the same shape and spread about the line.
- The residual errors are mutually independent (no pattern).
- The data are produced from a well-designed, random sample or randomized experiment.



The y values for each x value are normally distributed about the line with the same standard deviation. For each x value, the mean of the y values lies on the regression line. More y values lie near the line than are scattered further away from the line.

Section Review

Linear regression is a procedure for fitting a straight line of the form $\hat{y} = a + bx$ to data. The conditions for regression are:

- **Linear** In the population, there is a linear relationship that models the average value of y for different values of x.
- **Independent** The residuals are assumed to be independent.
- **Normal** The y values are distributed normally for any value of x.

- **Equal variance** The standard deviation of the y values is equal for each x value.
- **Random** The data are produced from a well-designed random sample or randomized experiment.

The slope b and intercept b of the least-squares line estimate the slope β and intercept α of the population (true) regression line. To estimate the population standard deviation of y, σ, use the standard deviation of the residuals, s. $s = \sqrt{\frac{SEE}{n-2}}$. The variable ρ (rho) is the population correlation coefficient. To test the null hypothesis H_0: ρ = *hypothesized value*, use a linear regression t-test. The most common null hypothesis is H_0: ρ = 0, which indicates there is no linear relationship between x and y in the population. The TI-83, 83+, 84, 84+ calculator function LinRegTTest can perform this test (STATS TESTS LinRegTTest).

Formula Review

Least Squares Line or Line of Best Fit:

$$\hat{y} = a + bx$$

where

a = y-coordinate of the y-intercept

b = slope

Standard deviation of the residuals:

$$s = \sqrt{\frac{SEE}{n-2}}$$

where

SSE = sum of squared errors

n = the number of data points

Exercise:

Problem:

When testing the significance of the correlation coefficient, what is the null hypothesis?

Exercise:

Problem:

When testing the significance of the correlation coefficient, what is the alternative hypothesis?

Solution:

H_a: $\rho \neq 0$

Exercise:

Problem:

If the level of significance is 0.05 and the p-value is 0.04, what conclusion can you draw?

Homework

Exercise:

Problem:

If the level of significance is 0.05 and the p-value is 0.06, what conclusion can you draw?

Solution:

We do not reject the null hypothesis. There is not sufficient evidence to conclude that there is a significant linear relationship between x and y because the correlation coefficient is not significantly different from zero.

Exercise:

Problem:

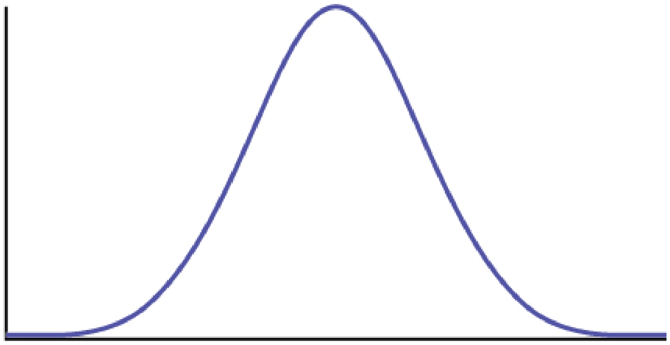
If there are 15 data points in a set of data, what is the number of degree of freedom?

Solutions Sheets

Hypothesis Testing with One Sample

Class Time: ___________________________
Name: _______________________________________

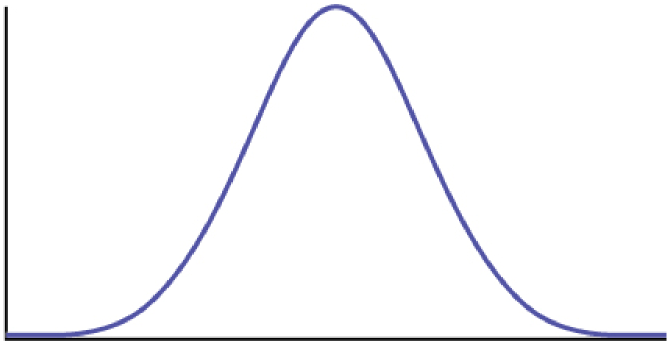
a. H_0 : ________
b. H_a : ________
c. In words, **CLEARLY** state what your random variable X or $\widehat{P}$ represents.
d. State the distribution to use for the test.
e. What is the test statistic?
f. What is the p-value? In one or two complete sentences, explain what the p-value means for this problem.
g. Use the previous information to sketch a picture of this situation. CLEARLY, label and scale the horizontal axis and shade the region(s) corresponding to the p-value.



h. Indicate the correct decision ("reject" or "do not reject" the null hypothesis), the reason for it, and write an appropriate conclusion, using **complete sentences**.

   i. Alpha: ________
   ii. Decision: ________
   iii. Reason for decision: ________
   iv. Conclusion: ________

i. Construct a 95% confidence interval for the true mean or proportion. Include a sketch of the graph of the situation. Label the point estimate and the lower and upper bounds of the confidence interval.

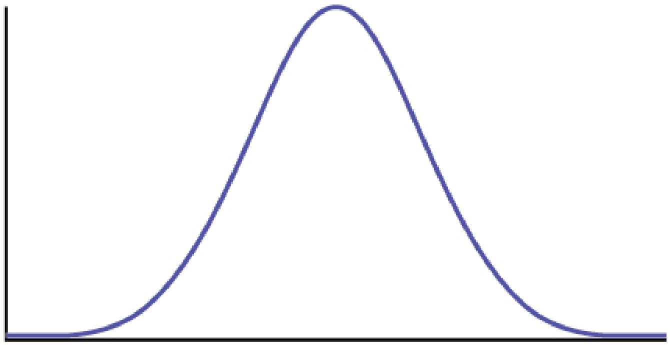


Hypothesis Testing with Two Samples

Class Time: ___________________________

Name: _______________________________________

a. H_0 : ________
b. H_a : ________
c. In words, **clearly** state what your random variable X X , $\hat{P}$ $\hat{P}$, or X_d represents.
d. State the distribution to use for the test.
e. What is the test statistic?
f. What is the *p*-value? In one to two complete sentences, explain what the *p*-value means for this problem.
g. Use the previous information to sketch a picture of this situation. **CLEARLY** label and scale the horizontal axis and shade the region(s) corresponding to the *p*-value.



h. Indicate the correct decision ("reject" or "do not reject" the null hypothesis), the reason for it, and write an appropriate conclusion, using **complete sentences**.

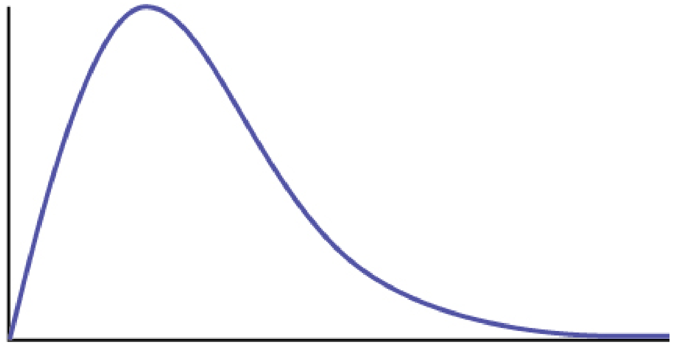
  a. Alpha: ________
  b. Decision: ________
  c. Reason for decision: ________
  d. Conclusion: ________

i. In complete sentences, explain how you determined which distribution to use.

The Chi-Square Distribution

Class Time: ______________________________
Name: ____

a. H_0 : ________
b. H_a : ________
c. What are the degrees of freedom?
d. State the distribution to use for the test.
e. What is the test statistic?
f. What is the *p*-value? In one to two complete sentences, explain what the *p*-value means for this problem.
g. Use the previous information to sketch a picture of this situation. **Clearly** label and scale the horizontal axis and shade the region(s) corresponding to the *p*-value.



h. Indicate the correct decision ("reject" or "do not reject" the null hypothesis) and write appropriate conclusions, using **complete sentences.**

    i. Alpha: ________
    ii. Decision: ________
    iii. Reason for decision: ________
    iv. Conclusion: ________

Tables

This module contains links to government site tables used in statistics.

Tables (NIST/SEMATECH e-Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook/, January 3, 2009)

- Student *t* table
- Normal table
- Chi-Square table
- F-table
- All four tables can be accessed by going to

95% Critical Values of the Sample Correlation Coefficient Table

- 95% Critical Values of the Sample Correlation Coefficient

Data Sets

Lap Times

The following tables provide lap times from Terri Vogel's log book. Times are recorded in seconds for 2.5-mile laps completed in a series of races and practice runs.

	Lap 1	**Lap 2**	**Lap 3**	**Lap 4**	**Lap 5**	**Lap 6**	**Lap 7**
Race 1	135	130	131	132	130	131	133
Race 2	134	131	131	129	128	128	129
Race 3	129	128	127	127	130	127	129
Race 4	125	125	126	125	124	125	125
Race 5	133	132	132	132	131	130	132
Race 6	130	130	130	129	129	130	129
Race 7	132	131	133	131	134	134	131
Race 8	127	128	127	130	128	126	128

	Lap 1	**Lap 2**	**Lap 3**	**Lap 4**	**Lap 5**	**Lap 6**	**Lap 7**
Race 9	132	130	127	128	126	127	124
Race 10	135	131	131	132	130	131	130
Race 11	132	131	132	131	130	129	129
Race 12	134	130	130	130	131	130	130
Race 13	128	127	128	128	128	129	128
Race 14	132	131	131	131	132	130	130
Race 15	136	129	129	129	129	129	129
Race 16	129	129	129	128	128	129	129
Race 17	134	131	132	131	132	132	132
Race 18	129	129	130	130	133	133	127
Race 19	130	129	129	129	129	129	128
Race 20	131	128	130	128	129	130	130

Race Lap Times (in seconds)

	Lap 1	**Lap 2**	**Lap 3**	**Lap 4**	**Lap 5**	**Lap 6**	**Lap 7**
Practice 1	142	143	180	137	134	134	172
Practice 2	140	135	134	133	128	128	131
Practice 3	130	133	130	128	135	133	133
Practice 4	141	136	137	136	136	136	145
Practice 5	140	138	136	137	135	134	134
Practice 6	142	142	139	138	129	129	127
Practice 7	139	137	135	135	137	134	135
Practice 8	143	136	134	133	134	133	132
Practice 9	135	134	133	133	132	132	133
Practice 10	131	130	128	129	127	128	127

	Lap 1	**Lap 2**	**Lap 3**	**Lap 4**	**Lap 5**	**Lap 6**	**Lap 7**
Practice 11	143	139	139	138	138	137	138
Practice 12	132	133	131	129	128	127	126
Practice 13	149	144	144	139	138	138	137
Practice 14	133	132	137	133	134	130	131
Practice 15	138	136	133	133	132	131	131

Practice Lap Times (in seconds)

Stock Prices

The following table lists initial public offering (IPO) stock prices for all 1999 stocks that at least doubled in value during the first day of trading.

$17.00	$23.00	$14.00	$16.00	$12.00	$26.00
$20.00	$22.00	$14.00	$15.00	$22.00	$18.00
$18.00	$21.00	$21.00	$19.00	$15.00	$21.00
$18.00	$17.00	$15.00	$25.00	$14.00	$30.00
$16.00	$10.00	$20.00	$12.00	$16.00	$17.44

$16.00	$14.00	$15.00	$20.00	$20.00	$16.00
$17.00	$16.00	$15.00	$15.00	$19.00	$48.00
$16.00	$18.00	$9.00	$18.00	$18.00	$20.00
$8.00	$20.00	$17.00	$14.00	$11.00	$16.00
$19.00	$15.00	$21.00	$12.00	$8.00	$16.00
$13.00	$14.00	$15.00	$14.00	$13.41	$28.00
$21.00	$17.00	$28.00	$17.00	$19.00	$16.00
$17.00	$19.00	$18.00	$17.00	$15.00	
$14.00	$21.00	$12.00	$18.00	$24.00	
$15.00	$23.00	$14.00	$16.00	$12.00	
$24.00	$20.00	$14.00	$14.00	$15.00	
$14.00	$19.00	$16.00	$38.00	$20.00	
$24.00	$16.00	$8.00	$18.00	$17.00	
$16.00	$15.00	$7.00	$19.00	$12.00	
$8.00	$23.00	$12.00	$18.00	$20.00	
$21.00	$34.00	$16.00	$26.00	$14.00	

IPO Offer Prices

References

Data compiled by Jay R. Ritter of University of Florida using data from *Securities Data Co.* and *Bloomberg*.

Notes for the TI-83, 83+, 84, and 84+ Calculators

Quick Tips

Legend

- ▭

  represents a button press
- `[ ]` represents yellow command or green letter behind a key
- `< >` represents items on the screen

To adjust the contrast
Press

2nd

, then hold

▲

to increase the contrast or

▼

to decrease the contrast.

To capitalize letters and words
Press

ALPHA

to get one capital letter, or press

2nd

, then

ALPHA

to set all button presses to capital letters. You can return to the top-level button values by pressing

ALPHA

again.

To correct a mistake
If you hit a wrong button, just hit

CLEAR

and start again.

To write in scientific notation
Numbers in scientific notation are expressed on the TI-83, 83+, 84, and 84+ using E notation, such that...

- 4.321 E 4 = 4.321×10^4
- 4.321 E –4 = 4.321×10^{-4}

To transfer programs or equations from one calculator to another:
Both calculators: Insert your respective end of the link cable cable and press

2nd

, then `[LINK]`.

Calculator receiving information:

Use the arrows to navigate to and select`<RECEIVE>`
Press ENTER.

Calculator sending information:

Press appropriate number or letter.
Use up and down arrows to access the appropriate item.

Press ENTER to select item to transfer.

Press right arrow to navigate to and select `<TRANSMIT>`.
Press ENTER .

Note:
Note
ERROR 35 LINK generally means that the cables have not been inserted far enough.

Both calculators: Insert your respective end of the link cable cable Both calculators: press

2nd

, then `[QUIT]` to exit when done.

Manipulating One-Variable Statistics

Note:
Note
These directions are for entering data with the built-in statistical program.

Data	**Frequency**

Data	**Frequency**
–2	10
–1	3
0	4
1	5
3	8

Sample Data We are manipulating one-variable statistics.

To begin:

1. Turn on the calculator.

   ON

2. Access statistics mode.

   STAT

3. Select `<4:ClrList>` to clear data from lists, if desired.

   4

   ,

   ENTER

4. Enter list `[L1]` to be cleared.

   2nd

   , `[L1]` ,

   ENTER

5. Display last instruction.

   2nd

   , `[ENTRY]`

6. Continue clearing remaining lists in the same fashion, if desired.

   ◀

   ,

   2nd

   , `[L2]` ,

   ENTER

7. Access statistics mode.

   STAT

8. Select `<1:Edit . . .>`

   ENTER

9. Enter data. Data values go into `[L1]`. (You may need to arrow over to `[L1]`).

   - Type in a data value and enter it. (For negative numbers, use the negate (-) key at the bottom of the keypad).

     (−)

     ,

     9

     ,

     ENTER

  - Continue in the same manner until all data values are entered.

10. In `[L2]`, enter the frequencies for each data value in `[L1]`.

    - Type in a frequency and enter it. (If a data value appears only once, the frequency is "1").

      4

      ,

      ENTER

    - Continue in the same manner until all data values are entered.

11. Access statistics mode.

    STAT

12. Navigate to `<CALC>`.

13. Access `<1:1-var Stats>`.

    ENTER

14. Indicate that the data is in `[L1]`...

    2nd

    , `[L1]` ,

    ,

15. ...and indicate that the frequencies are in `[L2]`.

    2nd

    , `[L2]` ,

    ENTER

16. The statistics should be displayed. You may arrow down to get remaining statistics. Repeat as necessary.

Drawing Histograms

Note:
Note
We will assume that the data is already entered.

We will construct two histograms with the built-in STATPLOT application. The first way will use the default ZOOM. The second way will involve customizing a new graph.

1. Access graphing mode.

   2nd

   , `[STAT PLOT]`

2. Select `<1:plot 1>` to access plotting - first graph.

   ENTER

3. Use the arrows navigate go to `<ON>` to turn on Plot 1.
   `<ON>` ,

   ENTER

4. Use the arrows to go to the histogram picture and select the histogram.

   ENTER

5. Use the arrows to navigate to `<Xlist>`.

6. If "L1" is not selected, select it.

   2nd

   , `[L1]` ,

   ENTER

7. Use the arrows to navigate to `<Freq>`.

8. Assign the frequencies to `[L2]`.

   2nd

   , `[L2]` ,

   ENTER

9. Go back to access other graphs.

   2nd

   , `[STAT PLOT]`
10. Use the arrows to turn off the remaining plots.
11. **Be sure to deselect or clear all equations before graphing.**

To deselect equations:

1. Access the list of equations.

   Y=

2. Select each equal sign (=).

   ▼

   ▶

   ENTER

3. Continue, until all equations are deselected.

To clear equations:

1. Access the list of equations.

   Y=

2. Use the arrow keys to navigate to the right of each equal sign (=) and clear them.

   ▼

   ▶

   CLEAR

3. Repeat until all equations are deleted.

To draw default histogram:

1. Access the ZOOM menu.

   ZOOM

2. Select `<9:ZoomStat>`.

   9

3. The histogram will show with a window automatically set.

To draw custom histogram:

1. Access window mode to set the graph parameters.

   WINDOW

2.
   - $X_{\min} = -2.5$
   - $X_{\max} = 3.5$
   - $X_{scl} = 1$ (width of bars)
   - $Y_{\min} = 0$
   - $Y_{\max} = 10$
   - $Y_{scl} = 1$ (spacing of tick marks on *y*-axis)
   - $X_{res} = 1$

3. Access graphing mode to see the histogram.

   GRAPH

To draw box plots:

1. Access graphing mode.

   2nd

   , `[STAT PLOT]`

2. Select `<1:Plot 1>` to access the first graph.

   ENTER

3. Use the arrows to select `<ON>` and turn on Plot 1.

   ENTER

4. Use the arrows to select the box plot picture and enable it.

   ENTER

5. Use the arrows to navigate to `<Xlist>`.

6. If "L1" is not selected, select it.

   2nd

   , `[L1]` ,

   ENTER

7. Use the arrows to navigate to `<Freq>`.

8. Indicate that the frequencies are in `[L2]`.

   2nd

   , `[L2]` ,

ENTER

9. Go back to access other graphs.

   2nd

   , `[STAT PLOT]`
10. **Be sure to deselect or clear all equations before graphing** using the method mentioned above.
11. View the box plot.

    GRAPH

    , `[STAT PLOT]`

Linear Regression

Sample Data

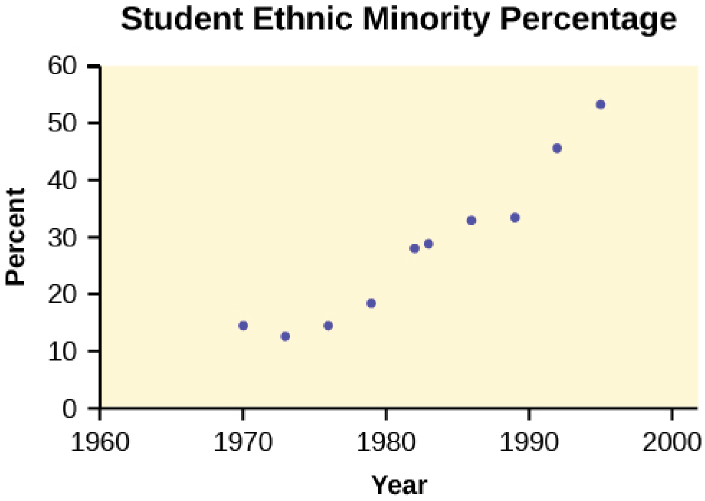
The following data is real. The percent of declared ethnic minority students at De Anza College for selected years from 1970–1995 was:

Year	Student Ethnic Minority Percentage
1970	14.13
1973	12.27
1976	14.08
1979	18.16

Year	**Student Ethnic Minority Percentage**
1982	27.64
1983	28.72
1986	31.86
1989	33.14
1992	45.37
1995	53.1

The independent variable is "Year," while the independent variable is "Student Ethnic Minority Percent."

Student Ethnic Minority Percentage



By hand, verify the scatterplot above.

Note:

> Note
> The TI-83 has a built-in linear regression feature, which allows the data to be edited.The *x*-values will be in `[L1]`; the *y*-values in `[L2]`.

To enter data and do linear regression:

1. ON Turns calculator on.

   ON

2. Before accessing this program, be sure to turn off all plots.
   - Access graphing mode.

     2nd

     , `[STAT PLOT]`
   - Turn off all plots.

     4

     ,

     ENTER
3. Round to three decimal places. To do so:
   - Access the mode menu.

     MODE

     , `[STAT PLOT]`
   - Navigate to `<Float>` and then to the right to `<3>`.

     ▼

     ▶

    - All numbers will be rounded to three decimal places until changed.

      ENTER

4. Enter statistics mode and clear lists `[L1]` and `[L2]`, as describe previously.

   STAT

   ,

   4

5. Enter editing mode to insert values for *x* and *y*.

   STAT

   ,

   ENTER

6. Enter each value. Press

   ENTER

   to continue.

To display the correlation coefficient:

1. Access the catalog.

   2nd

   , `[CATALOG]`

2. Arrow down and select `<DiagnosticOn>`

   ▼

   … ,

ENTER

,

ENTER

3. r and r^2 will be displayed during regression calculations.

4. Access linear regression.

   STAT

   ▶

5. Select the form of *y* = *a* + *bx*.

   8

   ,

   ENTER

The display will show:

LinReg

- *y* = *a* + *bx*
- *a* = –3176.909
- *b* = 1.617
- *r* = 2 0.924
- *r* = 0.961

This means the Line of Best Fit (Least Squares Line) is:

- *y* = –3176.909 + 1.617*x*
- Percent = –3176.909 + 1.617 (year #)

The correlation coefficient *r* = 0.961

To see the scatter plot:

1. Access graphing mode.

   2nd

   , `[STAT PLOT]`

2. Select `<1:plot 1>` To access plotting - first graph.

   ENTER

3. Navigate and select `<ON>` to turn on Plot 1.
   `<ON>`

   ENTER

4. Navigate to the first picture.

5. Select the scatter plot.

   ENTER

6. Navigate to `<Xlist>`.
7. If `[L1]` is not selected, press

   2nd

   , `[L1]` to select it.

8. Confirm that the data values are in `[L1]`.
   `<ON>`

   ENTER

9. Navigate to `<Ylist>`.

10. Select that the frequencies are in `[L2]`.

    2nd

, `[L2]` ,

ENTER

11. Go back to access other graphs.

    2nd

    , `[STAT PLOT]`
12. Use the arrows to turn off the remaining plots.
13. Access window mode to set the graph parameters.

    WINDOW

    - $X_{\text{min}} = 1970$
    - $X_{\text{max}} = 2000$
    - $X_{scl} = 10$ (spacing of tick marks on *x*-axis)
    - $Y_{\text{min}} = -0.05$
    - $Y_{\text{max}} = 60$
    - $Y_{scl} = 10$ (spacing of tick marks on *y*-axis)
    - $X_{res} = 1$

14. Be sure to deselect or clear all equations before graphing, using the instructions above.
15. Press the graph button to see the scatter plot.

    GRAPH

To see the regression graph:

1. Access the equation menu. The regression equation will be put into Y1.

   Y=

2. Access the vars menu and navigate to `<5: Statistics>`.

   VARS

   ,

   5

3. Navigate to `<EQ>`.
4. `<1: RegEQ>` contains the regression equation which will be entered in Y1.

   ENTER

5. Press the graphing mode button. The regression line will be superimposed over the scatter plot.

   GRAPH

To see the residuals and use them to calculate the critical point for an outlier:

1. Access the list. RESID will be an item on the menu. Navigate to it.

   2nd

   , `[LIST]`, `<RESID>`

2. Confirm twice to view the list of residuals. Use the arrows to select them.

   ENTER

   ,

   ENTER

3. The critical point for an outlier is: $1.9V\frac{\text{SSE}}{n-2}$ where:
   - n = number of pairs of data
   - SSE = sum of the squared errors
   - $\sum \text{residual}^2$
4. Store the residuals in `[L3]`.

   STO►

,

2nd

, `[L3]` ,

ENTER

5. Calculate the $\frac{(\text{residual})^2}{n-2}$. Note that $n-2=8$

2nd

, `[L3]` ,

x^2

,

÷

,

8

6. Store this value in `[L4]`.

STO▶

,

2nd

, `[L4]` ,

ENTER

7. Calculate the critical value using the equation above.

1

,

.

,

9

,

X

,

2nd

, `[V]` ,

2nd

, `[LIST]`

▶

,

▶

,

5

,

2nd

, `[L4]` ,

)

,

)

,

ENTER

8. Verify that the calculator displays: 7.642669563. This is the critical value.
9. Compare the absolute value of each residual value in `[L3]` to 7.64. If the absolute value is greater than 7.64, then the (x, y) corresponding point is an outlier. In this case, none of the points is an outlier.

To obtain estimates of *y* for various *x*-values:
There are various ways to determine estimates for "*y*." One way is to substitute values for "*x*" in the equation. Another way is to use the

TRACE

on the graph of the regression line.

TI-83, 83+, 84, 84+ instructions for distributions and tests

Distributions

Access `DISTR` (for "Distributions").

For technical assistance, visit the Texas Instruments website at http://www.ti.com and enter your calculator model into the "search" box.

Binomial Distribution

- `binompdf(n,p,x)` corresponds to $P(X = x)$
- `binomcdf(n,p,x)` corresponds to $P(X \leq x)$
- To see a list of all probabilities for *x*: 0, 1, . . . , *n*, leave off the "`x`" parameter.

Poisson Distribution

- `poissonpdf(λ,x)` corresponds to $P(X = x)$
- `poissoncdf(λ,x)` corresponds to $P(X \leq x)$

Continuous Distributions (general)

- $-\infty$ uses the value –1EE99 for left bound
- ∞ uses the value 1EE99 for right bound

Normal Distribution

- `normalpdf(x,μ,σ)` yields a probability density function value (only useful to plot the normal curve, in which case "`x`" is the variable)
- `normalcdf(left bound, right bound, μ, σ)` corresponds to $P(\text{left bound} < X < \text{right bound})$
- `normalcdf(left bound, right bound)` corresponds to $P(\text{left bound} < Z < \text{right bound})$ – standard normal
- `invNorm(p,μ,σ)` yields the critical value, k: $P(X < k) = p$
- `invNorm(p)` yields the critical value, k: $P(Z < k) = p$ for the standard normal

Student's *t*-Distribution

- `tpdf(x,df)` yields the probability density function value (only useful to plot the student-*t* curve, in which case "`x`" is the variable)
- `tcdf(left bound, right bound, df)` corresponds to $P(\text{left bound} < t < \text{right bound})$

Chi-square Distribution

- `X²pdf(x,df)` yields the probability density function value (only useful to plot the chi^2 curve, in which case "`x`" is the variable)
- `X²cdf(left bound, right bound, df)` corresponds to $P(\text{left bound} < X^2 < \text{right bound})$

F Distribution

- `Fpdf(x,dfnum,dfdenom)` yields the probability density function value (only useful to plot the F curve, in which case "`x`" is the variable)
- `Fcdf(left bound,right bound,dfnum,dfdenom)` corresponds to $P(\text{left bound} < F < \text{right bound})$

Tests and Confidence Intervals

Access `STAT` and `TESTS`.

For the confidence intervals and hypothesis tests, you may enter the data into the appropriate lists and press `DATA` to have the calculator find the sample means and standard deviations. Or, you may enter the sample means and sample standard deviations directly by pressing `STAT` once in the appropriate tests.

Confidence Intervals

- `ZInterval` is the confidence interval for mean when σ is known.
- `TInterval` is the confidence interval for mean when σ is unknown; *s* estimates σ.
- `1-PropZInt` is the confidence interval for proportion.

Note:
Note
The confidence levels should be given as percents (ex. enter "`95`" or "`.95`" for a 95% confidence level).

Hypothesis Tests

- `Z-Test` is the hypothesis test for single mean when σ is known.
- `T-Test` is the hypothesis test for single mean when σ is unknown; *s* estimates σ.
- `2-SampZTest` is the hypothesis test for two independent means when both σ's are known.
- `2-SampTTest` is the hypothesis test for two independent means when both σ's are unknown.
- `1-PropZTest` is the hypothesis test for single proportion.
- `2-PropZTest` is the hypothesis test for two proportions.
- X^2`-Test` is the hypothesis test for independence.

- `X`^{2}`GOF-Test` is the hypothesis test for goodness-of-fit (TI-84+ only).
- `LinRegTTEST` is the hypothesis test for Linear Regression (TI-84+ only).

Note:

Note

Input the null hypothesis value in the row below "`Inpt`." For a test of a single mean, "`μ∅`" represents the null hypothesis. For a test of a single proportion, "`p∅`" represents the null hypothesis. Enter the alternate hypothesis on the bottom row.